INDEX OF

THE JOURNAL OF EXPERIMENTAL EDUCATION

A QUARTERLY

EXECUTIVE EDITORS

JOHN SCHMID, Department of Research Statistics and Methodology, Colorado State College, Greeley. Chairman

PHILIP LAMBERT, Instructional Research Laboratory, School of Education, The University of Wisconsin, Madison.

September 1968 - June 1969

VOLUME XXXVII

DEMBAR EDUCATIONAL RESEARCH SERVICES, INC.

MADISON, WISCONSIN

A

Academic Attitudes and Performance as a Function of Differential Grading Systems: An Evaluation of Princeton's Pass-Fail System, Marvin Karlins, Spring 1969, 38.
After a Significant F in Anova? Robert R. Smawley, Spring 1969, 75.
Analysis of Means Technique Theory and Application, The; Shawky F. Karas, Winter 1968, 65.
Analytical Approach to Research on Instructional Methods, An; Nathanial L. Gage, Fall 1968,119.
Aquino, Milagros, Ludwig Mosberg, and Marge Sharron. Reading Comprehension Difficulty as a Function of Content Area and Linguistic Complexity, Summer 1969, 1.
Assessment of Mathematical Concepts of Five-Year-Old Children, Anthony N. Schwartz, Spring 1969, 67.
Attitudes of Elementary School Teachers Toward Counseling and Guidance in the Elementary Schools, Frederick J. Axelberd, Spring 1969, 1.
Attitudes Toward Services of State Departments of Education, W. L. Bashaw, James B. Kenny, William Landrum, R. Robert Rentz, and Foster Watkins, Spring 1969, 8.
Axelberd, Frederick J. Attitudes of Elementary School Teachers Toward Counseling and Guidance in the Elementary Schools, Spring 1969, 1.

B

Baggaley, Andrew R. A Symbolic System For Classifying Multivariate Procedures, Spring 1969, 5.
Baker, Eva L. Effects on Student Achievement of Behavioral and Nonbehavioral Objectives, Summer 1969, 5.
Bashaw, W. L., James B. Kenny, William Landrum, R. Robert Rentz, and Foster Watkins. Attitudes of Elementary School Teachers Toward Counseling and Guidance in the Elementary Schools, Spring 1969, 8.
Bayesian Inference, Jon C. Marshall, Winter 1968,71.
Begle, E. G. Curriculum Research in Mathematics, Fall 1968, 44.
Biographical Factors and Teacher Effectiveness, John A. Williamson, Spring 1969, 85.
Borderline Student as a Member of a Separate School Group, The; Donald R. Hetzner, Winter 1968, 48.
Borg, Walter R., Warren Kallenbach, Merva Morris, and Allen Friebel. Videotape Feedback and Microteaching in a Teacher Training Model, Summer 1969, 9.
Bright, R. Louis and Hendrik D. Gideonse. Research and Development Strategies: The Current Scene, Fall 1968, 139.

C

Caplin, Morris D. The Relationship Between Self Concept and Academic Achievement, Spring 1969, 13.
Carlson, Jerry S. and Frank L. Ryan. Levels of Cognitive Functioning as Related to Anxiety, Summer 1969, 17.
Chinn, Philip C., Oliver P. Kolstoe. Concept-Switching Task Performance as a Function of Rewards, Summer 1969, 21.
Coefficient Alpha, Test Reliability, and Heterogeneity of Score Distribution, Donald W. Zimmerman and Graham J. Burkheimer, Winter 1968, 90.
College Student Ratings of Instructors, Nicholas F. Rayder, Winter 1968, 76.
Comprehension of Repeated Time-Compressed Recordings, Thomas G. Sticht, Summer 1969, 60.
Concept-Switching Task Performance as a Function of Rewards, Philip C. Chinn, Oliver P. Kolstoe, Summer 1969, 21.
Confidence Level and Significance Level: Semantic Confusion or Logical Fallacy, William M. Stallings and Sushila Singhal, Summer 1969, 57.
Counts, George. The Creation and Solution of Unique Statistical Problems for Each Student, Spring, 1969, 17.
Covington, Martin. Promoting Creative Thinking in the Classroom, Fall 1968, 22.
Creation and Solution of Unique Statistical Problems for Each Student, The; George Counts, Spring 1969, 17.
Curriculum Research and Development in English, Robert C. Pooley, Fall 1968, 56.
Curriculum Research in Mathematics, F. G. Begle, Fall 1969, 44.
Curriculum Research in Science, Arthur H. Livermore, Fall 1968, 49.

D

deCharms, Richard and Virginia Carpenter. Measuring Motivation in Culturally Disadvantaged School Children, Fall 1968, 31.
Deliberative Attitudes, IQ and Social Class, T. Bentley Edwards, Winter 1968, 7.
Developmental Processes in Thought, Kenneth Lovell, Fall 1968, 14.
Development of Productive Thinking Skills in Fifth-Grade Children, The; James L. Wardrop, and others, Summer 1969, 67.
Dixon, Paul W., Nobuko K. Fukuda, and Anne E. Berens. Teachers' Ratings, Sex, and SCAT as Predictors of Rank in High School Class, Spring 1969, 21.
Douglass, Harl R. Note on Statistical Differences, Spring 1969, 7.
Dwyer, Francis M. An Experiment in Visual Learning at the Eleventh-Grade Level, Winter 1968, 1.

E

Edwards, T. Bentley. Deliberative Attitudes, IQ and Social Class, Winter 1968, 7.
Effects of Computer Based Resource Units Upon Instructional Behavior, The; George S. Holden, Spring 1969, 27.
Effects on Student Achievement of Behavioral and Nonbehavioral Objectives, Eva L. Baker, Summer 1969, 5.
Elementary School Form of the Dogmatism Scale, An; Russell L. Figert, Jr., Fall 1968, 19.
Experiment in Visual Learning at the Eleventh-Grade Level, An; Francis M. Dwyer, Winter 1968, 1.

F

Flanders, Ned A. Interaction Analysis and Inservice Training, Fall 1968, 126.
Free Recall of Verbal Material in the Classroom as a Function of Method of Presentation and Method of Practice, James R. Whitman, Summer 1969, 78.
Figert, Russell L., Jr. An Elementary School Form of the Dogmatism Scale, Winter 1968, 19.

G

Gage, Nathanial L. An Analytical Approach to Research on Instructional Methods, Fall 1968, 119.
Generating a Projected Criterion of Graduate School Success via Normative Judgment Analysis,

Samuel R. Houston, Winter 1968, 53.

Gideonse, Hendrik D. An Output-Oriented Model of Research and Development and Its Relationship to Educational Improvement, Fall 1968, 157.

Gill, Newell T., Roy King and Ron G. Wilburn. A Helping Relationship Experience in Teacher Education, Winter 1968, 24.

Goralski, Patricia J. and Joyce M. Kerl. Kindergarten Teacher Aides and Reading Readiness, Winter 1968, 34.

Gupta, Ram K. Treatment Comparisons: Item Responses in Multi-Factor Repeated Measurements Designs, Summer 1969, 26.

H

Haralick, Joy Gold. Teacher Acceptance of Administrative Action, Winter 1968, 39.

Helping Relationship Experience in Teacher Education, A; Newell T. Gill, Roy King, and Ron G. Wilburn, Winter 1968, 24.

Holden, George S. The Effects of Computer Based Resource Units Upon Instructional Behavior, Spring 1969, 27.

Hopkins, Kenneth D. and E. George Sitkei. Predicting Grade One Reading Performance Intelligence VS. Reading Readiness Tests, Spring 1969, 31.

Houston, Samuel R. Generating a Projected Criterion of Graduate School Success via Normative Judgment Analysis, Winter 1968, 53.

Hughes, Thomas M. A Study of the Relationship of Coping Strength to Self-Concept, School Achievement, and General Anxiety Level in Sixth Grade Pupils, Winter 1968, 59.

I

Increased Communication in Language Learning, Dieter M. Messner, Summer 1969, 45.

Instructional Research: Some Aspects of Its Status, Defects, and Needs, Arthur A. Lumsdaine, Fall 1968, 95.

Interaction Analysis and Inservice Training, Ned A. Flanders, Fall 1968, 126.

Instructional Television Around the World, Wilbur Schramm, Fall 1968, 89.

Is the Phonetic Generalization Hypothesis in Spelling Valid? Albert H. Yee, Summer 1969, 82.

K

Kane, Robert B. Semantic Differential Factor Structure with Concepts and Subjects from Education, Spring 1969, 34.

Karas, Shawky F. The Analysis of Means Technique Theory and Application, Winter 1968, 65.

Karlins, Marvin. Academic Attitudes and Performance as a Function of Differential Grading Systems: An Evaluation of Princeton's Pass-Fail System, Spring 1969, 38.

Kindergarten Teacher Aides and Reading Readiness, Patricia J. Goralski and Joyce M. Kerl, Winter 1968, 34.

Klausmeier, Herbert J. The Wisconsin Research and Development Center for Cognitive Learning, Fall 1968, 146.

L

Larsen, Suzanne W. and James R. Montgomery. A Study of Cluster Grouping, Summer 1969, 30.

Learning and Varying Sequences of Ninth-Grade Mathematics Materials, Fred Niedermeyer, Jerry Brown, and Bob Sulzen, Spring 1969, 61.

Levels of Cognitive Functioning as Related to Anxiety, Jerry S. Carlson and Frank L. Ryan, Summer 1969, 17.

Livermore, Arthur H. Curriculum Research in Science, Fall 1968, 49.

Logical Consistency of Student Teachers in Different Educational Environments, Henry R. Weinstock, Charles M. Peccolo, Floyd L. Coppedge, and Lloyd L. Coppedge, Spring 1969, 78.

Longitudinal Relationships Between Cognitive and Affective Learning Outcomes, Charles O. Neidt and Dalva E. Hedlund, Spring 1969, 56.

Lovell, Kenneth. Developmental Processes in Thought, Fall 1968, 14.

Lumsdaine, Arthur A. Instructional Research: Some Aspects of Its Status, Defects, and Needs, Fall 1968, 95.

M

Margules, Morton and Virginia M. Keith. A Study of Supervisors' Ratings of Most Effective and Least Effective Industrial Arts Teachers on Three Competency Dimensions, Summer 1969, 37.

Marshall, Jon C. Bayesian Inference, Winter 1968, 71.

Measuring Motivation in Culturally Disadvantaged School Children, Richard deCharms and Virginia Carpenter, Fall 1968, 31.

Messner, Dieter M. Increased Communication in Language Learning, Summer 1969, 45.

Movesian, George. Teacher Planning of Instructional Materials for Lower-Class Children, Summer 1969, 48.

Musella, Donald. Perceptual-Cognitive Style as Related to Self-Evaluation and Supervisor Rating by Student Teachers, Spring 1969, 51.

N

Neidt, Charles O. and Dalva E. Hedlund. Longitudinal Relationships Between Cognitive and Affective Learning Outcomes, Spring 1969, 56.

Niedermeyer, Fred, Jerry Brown, and Bob Sulzen. Learning and Varying Sequences of Ninth-Grade Mathematical Materials, Spring 1969, 61.

Note on Statistical Differences, Harl R. Douglass, Spring 1969, 7.

O

Organized Retention and Cued Recall, Endel Tulving, Fall, 1968, 3.

Otto, Wayne, Thomas C. Barrett and Theodore L. Harris. Research in Reading, Fall 1968, 65.

Output-Oriented Model of Research and Development and Its Relationship to Educational Improvement, Hendrik D. Gideonse, Fall 1968, 157.

P

Perceptual-Cognitive Style as Related to Self-Evaluation and Supervisor Rating by Student Teachers, Donald Musella, Spring 1969, 51.

Pooley, Robert C. Curriculum Research and Development in English, Fall 1968, 56.

Predicting Grade One Reading Performance: Intelligence VS. Reading Readiness Tests, Kenneth D. Hopkins and E. George Sitkei, Spring 1969, 31.

Prediction of Commitment to the Chemical Bond Approach to Teaching, Cynthia V. L. Ward, Winter 1968, 82.

Promoting Creative Thinking in the Classroom, Martin V. Covington, Fall 1968, 22.

Q

Quality of Written Expression of Third-Grade Children Under Differing Time Limits, The; Mary Jo Woodfin, Spring 1969, 89.

R

Rayder, Nicholas F. College Student Ratings of Instructors, Winter 1968, 76.

Reading Comprehension Difficulty as a Function of Content Area and Linguistic Complexity, Milagros Aquino, Ludwig Mosberg, and Marge Sharron, Summer 1969, 1.

Relationship Between Self Concept and Academic Achievement, The; Morris D. Caplin, Spring 1969, 13.

Research and Development Strategies: The Current Scene, R. Louis Bright and Hendrik D. Gideonse, Fall 1968, 139.

Research in Reading, Wayne Otto, Thomas C. Barrett and Theodore L. Harris, Fall 1968, 65.

Rosskopf, Myron F. Strategies for Concept Attainment in Mathematics, Fall 1968, 78.

S

Sawyer, R. N. and L. W. Martin. Specialized Study Skills, Developmental-Reading Instruction and Counseling, Summer 1969, 52.

Schwartz, Anthony N. Assessment of Mathematical Concepts of Five-Year-Old Children, Spring 1969, 67.

Schramm, Wilber. Instructional Television Around the World, Fall 1968, 89.

Semantic Differential Factor Structure With Concepts and Subjects from Education, Robert B. Kane, Spring 1969, 34.

Smawley, Robert R. After a Significant F in Anova? Spring 1969, 75.

SOCRATES, A Computer-Based Instructional System in Theory and Research, Lawrence M. Stolurow, Fall 1968, 102.

Specialized Study Skills, Developmental-Reading Instruction and Counseling, R. N. Sawyer and L. W. Martin, Summer 1969, 52.

Stallings, William M. and Sushila Singhal. Confidence Level and Significance Level: Semantic Confusion of Logical Fallacy, Summer 1969, 57.

Sticht, Thomas G. Comprehension of Repeated Time-Compressed Recordings, Summer 1969, 60.

Stolurow, Lawrence M. SOCRATES, A Computer-Based Instructional System in Theory and Research, Fall 1968, 102.

Strategies for Concept Attainment in Mathematics, Myron F. Rosskopf, Fall 1968, 78.

Study of Cluster Grouping, A; Suzanne W. Larsen and James R. Montgomery, Summer 1969, 30.

Study of Supervisors' Ratings of Most Effective and Least Effective Industrial Arts Teachers on Three Competency Dimensions, A; Morton Margules and Virginia M. Keith, Summer 1969, 37.

Study of the Relationship of Coping Strength to Self-Concept, School Achievement, and General Anxiety Level in Sixth Grade Pupils, Thomas M. Hughes, Winter 1968, 59.

Symbolic System for Classifying Multivariate Procedures, A; Andrew R. Baggaley, Spring, 1969, 5.

T

Teacher Acceptance of Administrative Action, Joy Gold Haralick, Winter 1968, 39.

Teacher Planning of Instructional Materials for Lower-Class Children, George Movesian, Summer 1969, 48.

Teachers' Ratings, Sex, and SCAT as Predictors of Rank in High School Class, Paul W. Dixon, Nobuko K. Fukuda, and Anne E. Berens, Spring 1969, 21.

Test Reliability and Parameters of Observed Score Distributions, Donald W. Zimmerman, Spring 1969, 92.

Titus, H. Edwin. The Use of Peer Nominations as a Predictor of Academic Success in College, Summer 1969, 63.

Treatment Comparisons: Item Responses in Multi-Factor Repeated Measurements Designs, Ram K. Gupta, Summer 1969, 26.

Tulving, Endel. Organized Retention and Cued Re-Call, Fall 1968, 3.

U

Use of Markov Chains to Estimate Effects of Feedback on Decision-Making, Gordon Welty, Summer 1969, 92.

Use of Peer Nominations as a Predictor of Academic Success in College, The; H. Edwin Titus, Summer 1969, 63.

V

Videotape Feedback and Microteaching in a Teacher Training Model, Walter R. Borg, Warren Kallenbach, Merva Morris, and Allen Friebel, Summer 1969, 9.

W

Ward, Cynthia V. L. Prediction of Commitment to the Chemical Bond Approach to Teaching, Winter 1968, 82.

Wardrop, James L., and others. The Development of Productive Thinking Skills in Fifth-Grade Children, Summer 1969, 67.

Weinstock, Henry R., Charles M. Peccolo, Floyd L. Coppedge, and Lloyd L. Coppedge, Logical Consistency of Student Teachers in Different Educational Environments, Spring 1969, 78.

Welty, Gordon. The Use of Experimental Designs in the Decision-Making Feedback Process, Winter 1968, 89.

Welty, Gordon. The Use of Markov Chains to Estimate Effects of Feedback on Decision-Making, Summer 1969, 92.

Whitman, James R. Free Recall of Verbal Material in the Classroom as a Function of Method of Presentation and Method of Practice, Summer 1969, 78.

Williamson, John A. Biographical Factors and Teacher Effectiveness, Spring 1969, 85.

Wisconsin Research and Development Center for Cognitive Learning, The; Herbert J. Klausmeier, Fall 1968, 146.

Woodfin, Mary Jo. The Quality of Written Expression of Third-Grade Children Under Differing Time Limits, Spring 1969, 89.

Y

Yee, Albert H. Is The Phonetic Generalization Hypothesis in Spelling Valid? Summer 1969, 82.

Z

Zimmerman, Donald W. and Graham J. Burkheimer. Coefficient Alpha, Test Reliability, and Heterogeneity of Score Distribution, Winter 1968, 90.

Zimmerman, Donald W. Test Reliability and Parameters of Observed Score Distribution, Spring 1969, 92.

Research and Development
Toward the Improvement of Education

Research and Development Toward the Improvement of Education[1]

Edited by

Herbert J. Klausmeier and George T. O'Hearn

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

DEMBAR EDUCATIONAL RESEARCH SERVICES, INC. Box 1148 Madison, Wisconsin 53701

Research and Development Toward the Improvement of Education

Published by Dembar Educational Research Services, Inc., Madison, Wisconsin.
Printed in the United States of America.

First Edition.

Library of Congress Catalog Card Number 68–58116

---

[1] The preparation of this report was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare, under the provisions of the Cooperative Research Program.

Center No. C-03/Contract OE 5-10-154

PREFACE

THE FIRST systematic attempt to apply research and development strategies to education occurred with the establishment of four research and development (R & D) centers in 1964 as part of the Cooperative Research Program of the United States Office of Education. The R & D Centers program attempted to combine the best features of the separate programs of the U.S.O.E.: basic and applied research, development, curriculum improvement, and demonstration. Thus, an R & D Center is organized to conduct basic research related to a significant problem area; to develop substantive and procedural products and conduct related research; to demonstrate the results of the research and development activities in a few school settings; and to take initiative for the eventual widespread dissemination of the research results and tested products. An R & D Center, located on a university campus, draws its professorial staff from the behavioral sciences, subject disciplines, communications, educational methodology, and other fields. It is a community of scholars, capable of dealing with a problem area effectively. In an R &D Center, then, human and financial resources are concentrated on a particular problem area over an extended period of time in order to secure a better understanding of that problem area and to improve related educational practices. Since 1964 the number of R & D Centers has increased from four to nine.

Our major universities, including the nine where R & D Centers are located, have a large pool of able researchers and many other resources. However, the development and application of research and development strategies to education is a pioneer venture. Developing imaginative strategies for extending knowledge about cognitive learning and improving related educational practices is a continuing concern of the Wisconsin Research and Development Center for Cognitive Learning.

In this Center, basic research is done to extend knowledge about the processes of cognitive learning, particularly those associated with concept learning and problem solving, and about the conditions associated with efficiency of learning in the cognitive domain. Some of this knowledge goes immediately into the development of instructional systems. A second objective of the Wisconsin R & D Center is the improvement of educational practices in the cognitive domain. The development and refinement of an instructional system related to a subject field, such as mathematics, requires appropriate strategies for dealing with the various components independently and also for interrelating them so that desired behavioral objectives are achieved. Three important components of the system are subject matter content and sequence; instructional materials and media; and instructional methods, including teacher behaviors and learner activities. A colloquium series was organized to bring information to the Center staff regarding these research and development strategies.

One part of this colloquium series was concerned with research on learning that has fairly direct implications for improving education. (A previous volume, Analyses of Concept Learning, Academic Press, 1966, on the other hand outlined more basic research.) The main portion of the colloquium series was addressed to development and related research dealing with subject matter content and sequence, instructional media, and instructional methods. This volume is organized in the same pattern.

As might be expected, the research methods reported by the participants are varied. In Part I, Endel Tulving of the University of Toronto shows that substantive results concerning information storage and retrieval, derived from the typical controlled laboratory experiment, require us to reconsider our notions about why persons do not effectively reproduce information they once learned. Kenneth Lovell, University of Leeds, shows that the charting of thought processes developmentally is moving ahead rapidly by means of Piaget's methods and refined variants thereof. Martin Covington, University of California, describes a straightforward attempt to develop materials designed to facilitate productive thinking in school children and to utilize the material in extending knowledge about children's productive thinking skills. Here there is a continuous interweaving of theorizing, developing material based on theory, and research activities concerning the abilities nurtured by the instructional program. Richard DeCharms and Virginia Carpenter of Washington University and Lindenwood College proceed in a similiar manner as Covington; however, their main concern is with motivational theory and the measurement of strength of motivation in culturally disadvantaged children. In these chapters substantive research results and theory properly receive most attention; however, sufficient attention is given to the procedures to make clear that extending knowledge about processes and conditions of learning can be accomplished by a variety of strategies.

In Part II, somewhat similar curriculum development and research procedures are outlined by the first three authors. E. G. Begle, Stanford University, outlines development and related research in mathematics, including his experiences with the widely known School Mathematics Study Group. Arthur Livermore, AAAS Commission on Science Education, briefly describes the large science curriculum projects, kindergarten to Grade 12, but gives major attention to the elementary program: Science—A Process Approach. Robert C. Pooley, The University of Wisconsin, also synthesizes many research and development activities related to the teaching of literature, composition, and language but gives particular attention to the Wisconsin English–Language Arts Project. Common to the curriculum development strategies in these subjects is an increasing emphasis on the structure of the discipline and a decreasing emphasis on other criteria such as social utility, difficulty, children's interests, and possible transfer to other subject fields. In the last two chapters of Part II, Wayne Otto, The University of Wisconsin, describes a rigorous and productive research effort in reading, and Myron Rosskopf, Teachers College, Columbia University, relates curriculum development in mathematics to recent research in concept learning.

In Part III, Wilbur Schramm, Stanford University, reports information on the use of instructional television. Three case studies are presented in sufficient detail to renew confidence in the value of careful observation in the natural environment by an interested, objective scholar. Lawrence Stolurow, Harvard University, describes the computer-based instructional system, SOCRATES, that he developed while at the University of Illinois. Problems associated with development of the hardware and software are noted briefly or may be inferred. The use of the computer in research to improve instruction and to validate behavior theory is noteworthy. Arthur Lumsdaine, University of Washington, deals with research and development in the entire field of instructional media.

In Part IV, N. L. Gage, Stanford University, outlines a recent concept, micro-teaching, and related research strategies. The concept promises to generate much research and development on instructional methods. In this approach, the teacher and student rather than the content, hardware, or program, are central in the educative process. Ned Flanders, University of Michigan, describes the new directions of his long-standing research on teacher-student interaction analysis. In the current model, the number of categories of interactions is reduced. By employing a computer for rapid feedback, the current model holds much promise of providing immediate (daily) correction of teacher–student interactions and also for supplying the extensive research data required in validating and extending the theoretical basis of the system of interaction analysis.

Part V of this volume contains papers which did not arise from the colloquium series. They are included because they contribute to the survey of research and development activities in education, particularly very recent trends. Louis Bright and Hendrick Gideonse of the U.S.O.E. outline the broad sweep of research and development strategies and activities, resulting from the passage of the Elementary and Secondary Education Act of 1965. Included here are the R & D centers, regional educational laboratories, Title III centers, and other R & D instrumentalities funded by other governmental agencies. The comprehensive educational research and development program of the U.S.O.E. merits careful attention by anyone interested in the improvement of education. A more complete description of one federally-supported R & D instrumentality, the Wisconsin Research and Development Center for Cognitive Learning, follows the chapter by Bright and Gideonse. Here we review the main components of a Center and their interrelations. Particular attention is given to the problem area of the Center; its programs, strategies, and outcomes; staffing pattern; administrative organization; locus of operation; and source and stability of funding.

In a new field, models are required as guides to operations. Many models have been outlined to indicate how systematic improvement in education might be managed through sequential research, development, and dissemination. None of the models thus far appeared particularly appropriate to the current scene. Hendrick Gideonse, active privately and not as a U.S.O.E. official, has been attempting for some time to formulate a comprehensive model that would simultaneously meet the test of reality and also provide a more adequate conceptualization of Research and Development activities designed to improve education. Fortunately his conceptualization was completed at the time the volume was in final editing. The Gideonse model is especially relevant to those who have been concerned with relating research, development, and educational improvement.

The colloquium series was planned and executed by the editors of this volume; Herbert J. Klausmeier assumed primary responsibility for organizing the colloquium series around topical areas of relevance to the R & D Center. He identified most of the participants and the broad organization framework for each participant's contribution. George T. O'Hearn participated in the planning and organization of the colloquium series, worked out the presentation schedule and other details with each participant, and handled editorial queries from the publisher. He also edited the papers and wrote the introductions to each part. Barbara Kennedy, Project Assistant, read the papers, checked the references, and put the references in APA style. The Wisconsin Research and Development Center for Cognitive Learning, the education research and development community, and the teaching profession generally is indebted to the internationally recognized scholars who made this volume possible.

Herbert J. Klausmeier, Professor of
Educational Psychology; Director,
Wisconsin Research and Development
Center for Cognitive Learning

George T. O'Hearn, Associate Professor,
Acting Director, Division of Education
University of Wisconsin—Green Bay

September 1968

CONTENTS

Preface

PART I Conditions and Processes of Learning

Introduction
George T. O'Hearn 1
University of Wisconsin–Green Bay

Chapter 1 Organized Retention and Cued Recall 3
Endel Tulving
University of Toronto, Toronto, Ontario

Chapter 2 Developmental Processes in Thought 14
Kenneth Lovell
University of Leeds, Leeds, England

Chapter 3 Promoting Creative Thinking in the Classroom 22
Martin V. Covington
University of California, Berkeley

Chapter 4 Measuring Motivation in Culturally Disadvantaged School Children 31
Richard deCharms
Washington University, St. Louis, Missouri
Virginia Carpenter
Lindenwood College, St. Charles, Missouri

PART II Subject Matter Content and Sequence

Introduction
George T. O'Hearn 42

Chapter 5 Curriculum Research in Mathematics 44
E. G. Begle
Stanford University, Stanford, California

Chapter 6 Curriculum Research in Science 49
Arthur H. Livermore
American Association for Advancement of Science
Commission on Science Education, Washington, D. C.

Chapter 7 Curriculum Research and Development in English 56
Robert C. Pooley
The University of Wisconsin, Madison

Chapter 8 Research in Reading 65
Wayne Otto and Thomas C. Barrett
The University of Wisconsin, Madison
Theodore L. Harris
University of Puget Sound, Tacoma, Washington

Chapter 9 Strategies for Concept Attainment in Mathematics 78
Myron F. Rosskopf
Teachers College, Columbia University, New York, New York

PART III Instructional Materials and Media

Introduction 87
George T. O'Hearn

Chapter 10 Instructional Television Around the World 89
Wilbur Schramm
Stanford University, Stanford, California

Chapter 11 Instructional Research: Some Aspects of Its Status, Defects, and Needs 95
Arthur A. Lumsdaine
University of Washington, Seattle

Chapter 12 SOCRATES, A Computer-Based Instructional System in Theory and Research 102
Lawrence M. Stolurow
Harvard Computing Center, Cambridge, Massachusetts

PART IV Instructional Methods and Teacher Behaviors

Introduction 118
George T. O'Hearn

Chapter 13 An Analytical Approach to Research on Instructional Methods 119
Nathanial L. Gage
Stanford University, Palo Alto, California

Chapter 14 Interaction Analysis and Inservice Training 126
Ned A. Flanders
The University of Michigan, Ann Arbor

PART V The Current Scene

Introduction 134
George T. O'Hearn

Chapter 15 Research and Development Strategies: The Current Scene 139
R. Louis Bright and Hendrik D. Gideonse
United States Office of Education, Washington, D. C.

Chapter 16 The Wisconsin Research and Development Center for Cognitive Learning 146
Herbert J. Klausmeier
The University of Wisconsin, Madison

Chapter 17 An Output-Oriented Model of Research and Development and Its Relationship to Educational Improvement 157
Hendrik D. Gideonse
United States Office of Education, Washington, D. C.

PART I

Conditions and Processes of Learning

UNDERSTANDING learning phenomena and conditions associated with efficiency of learning is a foremost concern of psychologists throughout Western civilization. Increasingly sophisticated experiments are being executed in laboratory and non-laboratory settings. Representative of work in this vast domain, with implications for education, are the following:

a) controlled observations to infer thought processes.
b) controlled experimentation in the laboratory to extend our knowledge of human memory.
c) development of material and related research to nurture the creative abilities of school children.
d) development of dependent measures to extend a motivational theory to school settings.

Permanence of learning necessarily requires the storage and recall of information. In Chapter 1, Endel Tulving discusses various verbal learning and memory experiments conducted in the laboratory at the University of Toronto, and suggests that "the success of retention and recall is to a large extent determined by what the subject does with the material he is to memorize, by the methods he uses to organize the material, and by the cues that he uses to effect the recall of the material." The data reported by Tulving support his contention that the ability to recall is determined more by the limitation of the retrieval mechanism than by the limitation of the storage mechanism. He concludes:

> *. . . that understanding of many phenomena of verbal learning and memory would be enhanced if we adopted a basic orienting attitude. That a person's mnemonic performance depends to a very large extent on the method that he wittingly or unwittingly uses to organize the material. The important function of such organization is to facilitate retrieval of information from the memory storage.*

The inferring of thought processes has always been difficult but is essential to explaining changes that occur ontogenetically. In "Developmental Processes in Thought," Kenneth Lovell, University of Leeds, reports on the stages of human intellectual growth hypothesized by Jean Piaget and his followers. Lovell suggests that the schemas interiorized by the child undergo constant modification and are the basis and limitation of all subsequent learning. "The implication of Piaget's findings is that we need to begin to stimulate the intellectual growth of the child in the early weeks of life." Research reported by Lovell confirms the general classification of the developmental stages of thought hypothesized by Piaget, and also indicates the importance of previous learning and the cultural milieu on the child.

Martin Covington in his paper "Promoting Creative Thinking in the Classroom" states, "In order to take full advantage of the unrestrictive atmosphere the child must first come to understand what constitutes creative ideas in the given situation, and how he can achieve such ideas for himself . . . In order to think creatively the child must have at his disposal a repertoire of creative thinking skills. To name a few: the ability to recognize gaps in existing information; a facility for formulating relevant questions, and a facility to the demands of the task so one can adequately judge the suitability of the proposed ideas."

The General Problem Solving Program described by Covington represents an attempt to initiate creative thought in students within a school setting and within a curriculum framework. The objective is to teach a number of cognitive skills which are fundamental to creative thinking and then determine whether such generalized skills can be applied to a specific subject matter area.

Before specified learning can take place the learner must seek and accept the learning situation. The probability that an individual will seek or accept a new learning situation is referred to as motivation by Richard deCharms and Virginia Carpenter. When a tendency to avoid a learning situation increases, i.e.., when there is a lack of motivation, the problem becomes crucial. The discussion of this problem by deCharms and Carpenter is based on their work with inner-city youngsters. In their chapter "Measuring Motivation in Culturally Disadvantaged School Children" the authors propose a model of motivation in which a person perceived to be originating his own behavior is called an "Origin" while a person perceived to be behaving in a certain way because of an external "power source" is said to be acting like a "Pawn." "Pawn behavior implies powerlessness, but origin behavior implies freedom of choice, not necessarily the ability to wield power over others." The theoretical framework of the "OP" model is discussed in detail. The n achievement procedures developed by McClelland were used in measuring motivation in a sample of culturally disadvantaged school children. The data obtained tend to support the theoretical point of view expressed

Organized Retention and Cued Recall[1]

Endel Tulving
University of Toronto
Toronto, Ontario

EVER SINCE Hermann Ebbinghaus (1885) brought phenomena of memory under experimental scrutiny in the psychological laboratory and William James (1890) eloquently epitomized the distilled wisdom of other great thinkers on the subject, research on remembering and forgetting has been growing by leaps and bounds. Everyone agrees that this research has by now produced an immense quantity of empirical data, but few are willing to claim that it has created any genuine novel conceptual insights. More often than not, new empirical findings have generated more heat than light at the theoretical level. Thus, despite everything we believe we know about the role of a multitude of experimentally manipulable variables on verbal learning, retention, and recall, we seem to be no nearer to the understanding of the basic nature of memory than were psychologists and other writers of previous generations. If someone asked us today what do we know that is truly worth knowing and that was not known to Ebbinghaus and James, we would be hard pressed to provide a convincing answer.

Those who share this pessimistic assessment of the state of affairs in research on verbal learning and memory have been quick to suggest remedies, but their advice has been inconsistent. Some have argued that experimental conditions between laboratories be more rigidly standardized, others would like to see a greater diversity of methods. Some advise that only limited sets of carefully calibrated materials be used, others insist on a fuller variety of materials. Some are convinced that true understanding of memory can be achieved only through mammoth parametric studies, others are equally certain that at least for the time being the solution lies in simpler "miniature" experiments. I am listing these conflicting attitudes not to pass judgment on them, but rather as a background for my own favorite complaint about the general orientation of research in this broad field.

My complaint is that far too little attention has been devoted to the implications of the well-known fact that subjects in verbal learning and memory experiments are not passive organisms. Too many students of human learning and memory have predicated the interpretation of their data and the formulation of their theoretical statements on the basic assumption that in learning and remembering certain things happen to the person or certain events take place in his nervous system, and that these events are governed by nothing more than the present and past inputs. Eventually, of course, science may be able to interpret all phenomena of memory without recourse to a highly sophisticated, conscious, active, thinking organism, but at the present time nothing much seems to be gained, and a good deal may be lost, if we pretend that the articulate human subject is a passive recipient of external stimulation, and that acquisition, retention, recall, transfer, interference, etc., occur in a mental vacuum. If we know anything at all about what subjects do in verbal learning and memory experiments, we know that they use plans (Miller, Galanter, and Pribram, 1960), that they consciously and deliberately employ strategies (e.g., Pollack, Johnson, and Knaff, 1959) or mediational devices (e.g., Bugelski, 1962; Underwood and Schulz, 1960), or simply that they think (e.g., Mandler, 1965). There are some who would argue that these covert activities of the subject, and their behavioral manifestations, merely accompany processes of memory and that they play no significant role in determining the quantity and quality of a person's mnemonic performance, but such views are rapidly losing their appeal. A great deal of evidence accumulated in recent years, added to evidence resurrected from isolated and forgotten experiments from the past, points to the witting or unwitting use of plans, strategies, and mediational devices by the subject as the sine qua non of remembering. A fuller understanding of what the subject does when faced with the requirements of memory tasks of various kinds, therefore, seems indispensable for the comprehension of many, if not all, phenomena of memory. In all probability, an active search for such understanding will turn out to be more fruitful than the mere introduction of standardized conditions of experimentation, calibration of materials, or mammoth experiments.

In the present paper I will describe some recent work we have been doing at Toronto to probe into some of the long-standing problems of memory. Two questions that I have been particularly interested in are, first, why is repetition effective in increasing the amount of material the subject can recall, and secondly, how does retrieval of mnemonic information from the memory storage take place. I will obviously not be able to answer these questions, but I would like to offer some suggestions relevant to them. These suggestions are firmly anchored to the basic assumption that the success of retention and recall is to a large extent determined by what the subject does with the material he is to memorize, by the methods he uses to organize the material, and by the cues that he uses to effect the recall of retained material. The data I will report, I believe, will offer good support to this assumption.

FREE RECALL AND ORGANIZATION

The experimental data come from experiments in which the subject's task is to learn, that is to commit to memory, a set of discrete verbal items. Some of these experiments are concerned with free recall of items, while others constitute variations on the theme of free recall. In the free-recall experiment the subject is presented with a list of familiar items, such as meaningful words, and his task is to remember as many of these items as possible, in any order that they occur to him. Thus the free-recall paradigm constitutes a test for the acquisiton, retention, and utilization of retained information about the membership of otherwise familiar items in an unfamiliar set. Free recall as a laboratory task, like other paradigms of verbal learning and memory, involves mnemonic performance of the kind often required of individuals outside the laboratory. A teacher's knowledge of names of students in a class, a person's memory for the names of guests at a cocktail party he attended the night before, and a child's recollection of animals he saw on his visit to the zoo are based on retention of information of the kind that we investigate in free-recall experiments.

I have been particularly interested in free-recall phenomena for three reasons. First, the requirements imposed on the mnemonic information processing mechanisms in the free-recall task are necessarily present in any other paradigm. The subject cannot learn the order of a set of items in a serial learning task, or specific inter-item associations in a paired-associate task, unless he knows what items constitute the set. This means that the processes involved in free recall are in some ways more basic than those underlying many other laboratory tasks in memory, and that therefore the study of these processes constitutes a logical starting point for the study of memory. Secondly, free-recall phenomena were ignored by students of verbal learning and memory for a long time. As a consequence they have not yet been fitted into a relatively fixed theoretical framework as have been serial- and paired-associate learning phenomena. Thus, when we are faced with the problem of interpreting experimental findings in free recall, we are not hampered by extant theoretical notions. Thirdly, and most importantly, it is easier to find out in free-recall situations what the subject is doing when he is memorizing a set of items. The task set to the subject in serial- and paired-associate paradigms is too rigidly prescribed by the experimenter and all that can be observed at the behavioral level is how well the subject can meet the experimenter's requirements. In free recall, the subject is told to recall as many items as possible, but he is left free to recall the items in any order. Examination of the order in which he recalls the items provides considerable insight into the subject's covert activities in the process of memorization.

In multi-trial free-recall tasks, the same set of items is presented on a number of successive trials, the order of presented items usually varying from trial to trial (cf. Waugh, 1961), and after each input trial the subject recalls all items he can. When the number of items the subject recalls are plotted against trials, a typical negatively accelerated learning curve results (Miller and McGill, 1952; Murdock, 1960; Tulving, 1962a, 1964; Waugh, 1961). One of the questions I wish to raise is, why does such learning, i. e., improvement in trial-to-trial recall, occur. I have argued elsewhere (Tulving, 1962a, 1964) that learning in multi-trial free recall is a consequence of the fact that the subject organizes list-items into larger units of material, into subjective units or S-units. At any given time the subject can retrieve from memory only a limited number of units of a particular set of materials, but the larger these units, that is, the more individual items each unit contains, the larger the number of items recalled (Miller, 1956a, 1956b).

ORGANIZATION AND LEARNING

The original evidence for this contention was provided by the observations that the amount of organization subjects imposed on the material increased systematically over trials, accompanying the increase in the number of items recalled, and that when the amount of practice was held constant, there was a sizable correlation between recall scores and organization scores as subject variables. Such correlational evidence, of course, is not very compelling. One can justifiably argue that organization and recall are two parallel but otherwise unrelated manifestations of some common underlying mechanism. Some writers (e.g., Asch and Ebenholtz, 1962) have indeed claimed that free-recall learning occurs independently of any inter-item associations, being a consequence of greater availability of items resulting from repeated presentation of items. To the extent that the development of S-units reflects the development of specific inter-item associations, the denial of the primary role of inter-item associations in free-recall learning also constitutes a denial of the dependence of learning on organizational processes. How can we evaluate the hypothesized relation between organization and free-recall learning more directly than is possible on the basis of observed correlations between the two measures?

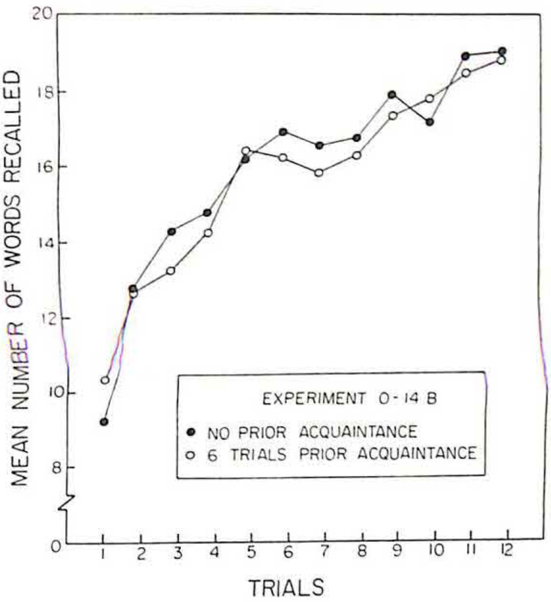
Two possibilities suggest themselves. The first is to present a set of items to the subject under conditions where he makes repeated responses to the items, but does not actively and deliberately try to organize the material into larger S-units. If learning depends upon organization, then such non-directed repetition should have little effect on recallability of items. The second possibility is to instruct subjects to use effective methods of organization. If learning depends on organization, such instructions would be expected to facilitate learning. We have done both types of experiments.

REPITITION AND ORGANIZATION

At first glance, it may seem difficult to create conditions under which subjects are repeatedly exposed to a set of items and where they make overt responses to these items, but do not think about and thus organize the items. Since we initially knew nothing about the problems involved, we decided to start with the most obvious approach, although, I must admit, at the time we did not have much faith in it. Subjects were simply shown a set of common words, asked to read these words aloud a number of times, and then, immediately after such reading, were given the same set of words to learn under the typical free-recall instructions (Tulving, 1966). A control group learned the same set of words under identical conditions, but without having seen the words beforehand.

The list to be learned contained twenty-two high-frequency English words. In the "reading task," the subjects in one group were shown each of the words six times, in an uninterrupted sequence of 132 exposures. The words were exposed one at a time by means of a memory drum, at the rate of one word/second, and the subjects were instructed to call out each word as it appeared. The subjects in the other group went through the same procedure with a list of twenty-two names. Immediately following the reading task, all subjects were asked to learn the same set of twenty-two words to which the first group had been exposed already. Typical free-recall instructions were given, words were presented at the rate of one word/second, and at the end of each input phase subjects had 60 seconds for oral recall. Every subject received twelve learning trials. Thus, both groups learned the identical list of words under identical conditions. The only difference was that the subjects in the Prior Acquaintance Group has just finished reading aloud the same words on six continuous reading "trials," while the subjects in the No Prior Acquaintance Group had been "warmed up" with a different set of words.

FIGURE 1. Learning curves for two groups of subjects. One group received six trials of prior acquaintance with the set of twenty-two words to be learned later, the other group had no such prior acquaintance.



The two learning curves from this experiment are shown in Figure 1. The Prior Acquaintance Group had a small, but statistically non-significant edge over the No Prior Acquaintance Group on the very first trial, but none thereafter. Over all twelve trials, the mean number of words recalled by the former group was 15.71, while the mean number of words recalled by the latter group was 15.91.

These results thus suggest that mere repetition of list items had no effect on subsequent memorization of these items. According to the present interpretation, repetition had no effect because the subjects did not attempt to organize items in any way. Post-experimental interviews revealed indeed that very few subjects had thought that they might be asked to recall the words they were reading in the original reading task. Most of them said they thought the experimenter was interested in how well they could read, how fast they could read or how clearly or consistently they could pronounce the words.

Although we did not give an immediate recall test to the subjects at the end of the preliminary reading task, it is extremely unlikely that they would not have remembered any words that they had seen. Thus, this experiment does not provide any evidence against incidental learning measured in a single recall test following a reading task of the kind we used. It only demonstrates that directed memorization under the conditions of free recall is not facilitated by whatever prior incidental learning has occurred. This is not a surprising result, if we assume that no organization occurs during the reading task, that incidental memory of items read, therefore, would involve individual items as S-units, and that the subject can only recall a constant limited number of S-units on any given trial in free recall. If the subject, on the first trial in the directed learning task, recalls some of the items he learned incidentally, he could not recall too many other items because his limited retrieval capacity would already be partially exhausted. One of the implications of these findings is that if the subject, before the presentation of a longer list for directed recall, were told what some of the items in the list are, his recall would be no greater than that of the subject who has no such prior information, provided the items are not meaningfully related to one another and that the rate of presentation is rather fast. The hypotheses that increased recall can occur only as a consequence of organization and that the retrieval capacity is limited to a fixed number of S-units would completely account for this finding, even though it could not be readily predicted from many other theoretical points of view.

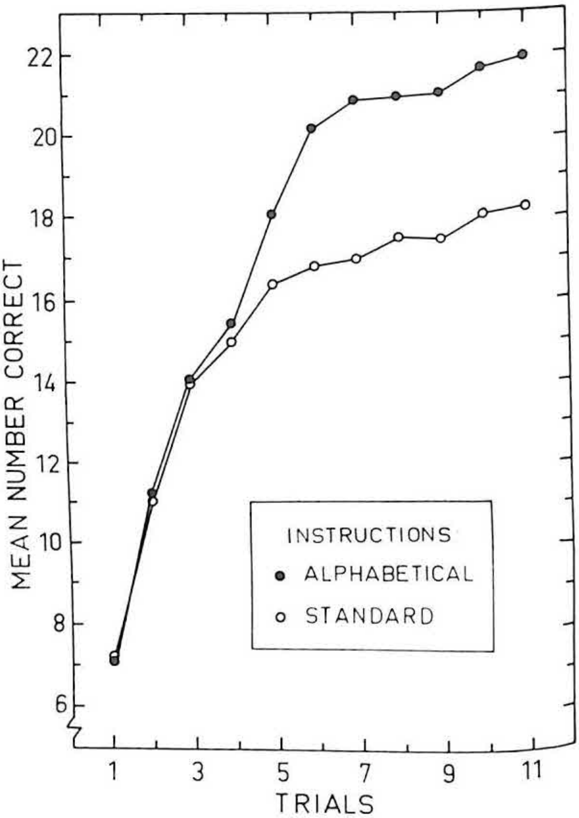
ALPHABETICAL ORGANIZATION

The second type of more direct test of the relation between learning and organization, as I said, involves manipulation of organizing strategies through instructions to subjects. An initial study of this kind was done as a classroom experiment (Tulving, 1962b) The subjects were eighty university students. They had to learn a list of twenty-two nouns, each beginning with a different letter, and the whole set covering a wide range of Thorndike-Lorge (1944) frequencies. In the input phase of each trial, words were projected on a screen at the rate of 1.5 second/word. In each output phase of 90 seconds duration subjects wrote down as many words as they remembered.

At the beginning of the experiment, all subjects were given typical free-recall instructions. For the first three trials all subjects worked under these identical instructions. Before the fourth trial, subjects were asked to open and read additional instructions on folded sheets handed out to them before the experiment. For half the subjects, the instructions simply reiterated what all subjects had been told at the beginning of the experiment, namely to "do your very best on each recall trial and put down as many words from the list as you can." The instructions for the other half of the subjects read: "Try to organize your recalled words alphabetically. When you look at the words on the screen, note their first letters, and make an attempt to associate the word with the letter. When you write the words down, go through the letters of the alphabet one at a time and try to remember the word that goes with each letter." Subjects in each group, the Standard Group and the Alphabetical Group, were given 30 seconds to read these additional instructions. The experiment was then continued under otherwise identical conditions for eight more trials.

The learning curves for the two groups are shown in Figure 2. For the first three trials, when both groups worked under identical standard instructions, the curves are almost indistinguishable indicating no differences in learning "ability" of the two groups.

FIGURE 2. Learning curves for two groups of subjects. After Trial 3, one group was given instructions to recall the words alphabetically, while the other group continued learning under standard instructions.



From Trial 4 on, however, the curves draw rapidly apart, the mean number of words recalled by the Alphabetical Group on Trial 5 equaling that of the Standard Group on the last trial of the experiment, Trial 11. Judging by the shape of the curve of the Standard Group, it appears that it would have taken them many additional trials to reach the level of performance comparable to that reached by the Alphabetical Group on Trial 6. The differences between the two groups would undoubtedly be even greater if it were not for the artificial limit on the performance of the Alphabetical subjects imposed by the "ceiling effect."

These results clearly indicate that methods of organization are very important in determining the course of memorization of a list of items. Alphabetical organization is apparently much more effective in this type of task than are other methods of organization – grouping of words by their meaning – used by the subjects in the Standard Group. The learning curve of the subjects in the Standard Group reflects their limited performance, but the limit is not wholly determined by their inadequate mnemonic ability. To a large extent, it is determined by their ignorance as to how to make best use of their mnemonic ability in this type of task. When subjects are given more powerful organizing instructions, their "ability" increases considerably.

One important implication of this observation – I am now going to digress somewhat from my main theme – has to do with the problem of individual differences in learning ability. It is commonplace to study individual differences in learning ability by dividing subjects into "fast" and "slow" learners on the basis of their performance in a learning task, and then look for similar differences between the two groups in various other situations, such as various psychometric tests. In the experiment I have just described, there were pronounced differences in the memorization performance between the subjects in two groups, but since the groups were selected randomly it is a foregone conclusion that they would have shown no differences on any other tests. How frequently are individual differences in learning "ability" determined by the same factors as those manipulated in the present experiment? The long history of failures to find powerful psychometric correlates of learning performance, or sizeable correlations between different learning tasks, suggests that individual differences in learning "ability" may be largely governed by individuals' knowledge of effective strategies of handling a particular task, rather than by the sheer goodness of their memory. At any rate, in the light of data from this experiment it seems that if we wish to add to our understanding of individual differences in learning "ability," a good place to start is to analyze the learners' strategies of memorizing in a given learning situation, and not to observe his behavior in some other extra-experimental situation.

ORGANIZATION AND LEARNING

Let me now return to the relation between organization and learning. We have seen that when a subject has to memorize a list of randomly selected words, alphabetical organization is more effective than other methods of organization. Alphabetical organization changes the task into a kind of paired associate learning task. The initial letter of each word serves as a cue for the recall of the whole word. The experimenter does not have to present these cues since an articulate subject can generate them on his own. We could say that in alphabetical organization each input word is cued with respect to its initial letter and that its retrieval is facilitated by the availability of relevant cues. When other methods of organization are used, such as grouping of items in terms of their associative meaning, their belongingness into superordinate conceptual categories, or the degree of their prior familiarity, input words are cued with respect to their properties other than the initial letter and their retrieval is presumably effected through these other cues. It is possible, of course, that a given word is cued in several ways and that the way it is cued changes in the course of the experiment, but under the usual time pressure in a typical experiment, multiple cueing is probably an exception rather than a rule.

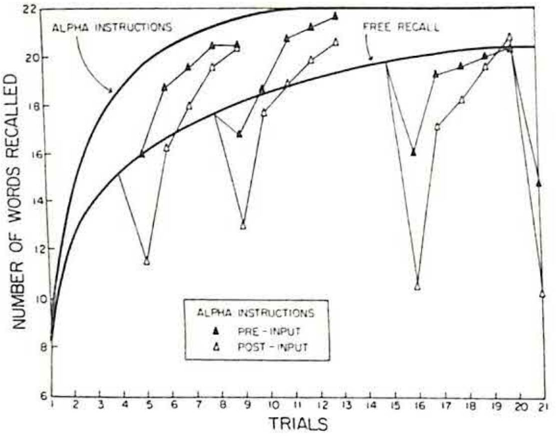
Suppose now that a subject has been exposed to a set of items for a number of trials, that he has not discovered alphabetical method of organization, and that he has been cueing the words in terms of their other characteristics. What would happen, if the subject were now told to switch to another method of organization, say alphabetical organization? We know already that alphabetical organization is likely to be more effective than other methods of organization and that therefore alerting the subject to it should facilitate learning. But it is also possible that changing the method of organization requires not only the adoption of the new strategy, but also the suppression of the previously used strategy. One might expect, therefore, that immediately after the introduction of alphabetical instructions the efficiency of the storing and retrieval mechanisms is temporarily reduced and that the disturbance is the greater the more highly developed the non-alphabetical organization.

Data relevant to these hypotheses come from an experiment done by Marcia Ozier as part of her PhD thesis (Ozier, 1965). Her subjects learned a list of twenty-two words, each beginning with a different letter. Words were projected on a screen at the rate of 1.5 second/word and subjects recorded their recall on paper-pullers that prevented them from seeing more than one word of their recall at any time. The duration of output phase of each trial was 90 seconds.

Most subjects were initially given standard free-recall instructions and then switched over to alphabetical instructions at some point in the experimental sequence. The two independent variables in the experiment were, first, the ordinal number of the "critical" trial on which alphabetical instructions were given, and second, pre-input versus post-input alphabetical instructions. Different groups of subjects were given alphabetical instructions on Trials 1, 5, 9, 16, and 21. In each of these groups, half the subjects were given alphabetical instructions before the input phase of the critical trial, while the other half received these instructions after the input phase, that is, before the output phase, of the critical trial. Before the critical trial, subjects in a given group worked under standard free-recall instructions. Thus, for instance, one group of subjects worked under free-recall instructions for eight trials, then received alphabetical instructions, followed by the input and output phases of the ninth and subsequent trials. Another group worked under standard free-recall instructions for eight trials, was then shown the words in the input phase of the ninth trial, and then was asked to start recalling words in the alphabetical order, beginning with the immediately following output phase. There was thus a total of ten independent groups, with twelve subjects per group. Scoring was lenient; even if the subject had deviated somewhat from the alphabetical order when he was supposed to recall the words alphabetically, he was given credit for each word correctly recalled. Subjects who obviously did not follow alphabetical instructions, however, were replaced.

The somewhat schematized results of this experiment are shown in Figure 3. The upper heavy smoothed curve shows the learning curve for the subjects who received alphabetical instructions before Trial 1, the lower heavy smoothed curve represents the free-recall conditions. Subjects who had received pre-input alphabetical instructions on Trial 1 had a mean recall of approximately eight words on that trial, while those who had received

FIGURE 3. The effect of alphabetical organizing instructions given at various points in the course of learning. See text for explanation.



post-input alphabetical instructions on Trial 1 recalled approximately five words, a difference of three words. The learning curve for the latter group is not shown on the graph. It reached the learning curve of the former group by the fifth trial, being practically identical with it thereafter.

The other data in Figure 3 show the initial drop and subsequent rapid recovery in recall of groups who were given alphabetical instructions at Trials 5, 9, 16, and 21. For each of these groups, with the exception of those for whom Trial 21 was the critical trial, mean recall scores are shown for five trials after the critical trial. These data show, first, that immediately after the introduction of alphabetical instructions there is a drop in performance – with the exception of pre-input instructions on Trial 5 – with the magnitude of the drop varying directly with the ordinal number of the critical trial and being consistently greater for post-input alphabetical instructions, and second, that the subsequent recovery of the effectiveness of alphabetical instructions given on the critical trial varies inversely with the ordinal number of the critical trial.

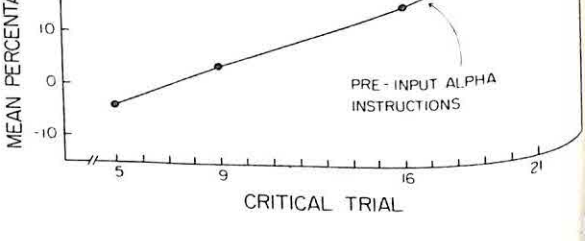
The percentage loss in recall in the first output phase following the introduction of alphabetical instructions is plotted as a function of the ordinal number of the critical trial, for both pre-input and post-input instructions, in Figure 4. Each data point in this graph represents the mean difference in recall between the number of words recalled on the critical trial and the immediately preceding trial divided by the mean number of words recalled on the immediate pre-critical trial. The data plotted in Figure 3 showed that the drop in performance after the introduction of alphabetical instructions, measured in terms of absolute number of items, varied directly with the ordinal number of the critical trial. The data shown in Figure 4 indicate that the same relation holds between percentage loss in recall and the ordinal number of the critical trial.

From previous experiments we already know that organization of material increases over trials of practice in the free-recall situation (Bousfield, Puff, and Cowan, 1964; Ehrlich, 1965; Tulving, 1962a, 1964). Therefore, manipulating the temporal point at which subjects are switched from other forms of organization to alphabetical organization essentially amounts to manipulation of the degree of existing organization that has to be changed. The results of Ozier's experiment suggest that the stronger the pre-existing organization, that is, the more firmly the list-items are cued with respect to characteristics of words other than their initial letters, the more disruptive is the requirement that the subject change his system of cueing and the longer it takes for the otherwise powerful alphabetical system of organization to become effective.

AVAILABILITY AND ACCESSIBILITY OF STORED INFORMATION

I have so far talked about two general classes of methods of organization: organization of words into larger groups in terms of their perceived, discovered, or fabricated meaningful relations to one another; and organization in terms of structural features of words, such as initial letters. In either case it is as if words stored in memory – or information about words to be reproduced – are in storage in a way that would facilitate retrieval: two or more related words are stored in or near the same location, such that when the retrieval mechanism finds one, the other one is also easily located, or else a word is stored in a location in which its initial letter is stored. Thus, one part of the stored information can be thought of as a retrieval cue for another part. Organization of information in the storage, however, can facilitate retrieval only to the extent that retrieval cues are available at the time of recall. Alphabetical organization is more effective than organization of items in terms of their semantic

FIGURE 4. Percentage loss of recall on the trial immediately following the introduction of alphabetical organizing instructions, as a function of the "critical" trial on which alphabetical instructions were given.



features because the initial letters of the words, serving as retrieval cues are readily available to the subject at the time of recall.

This kind of speculation implies that organization of material in the memory storage, or organized retention, does not necessarily increase the storage capacity, but rather makes the stored information more accessible to recall by providing distinctive retrieval cues through which individual items of information can be reached. It also implies that the amount of information potentially available in the storage at a given time may be greater than is apparent in typical unaided recall. In other words, limitation of memory is primarily determined by the limited capacity of the retrieval mechanism and not necessarily by the limited capacity of the storage mechanism.

In experiments of the kind that we have examined so far it is impossible to distinguish between retention and recall. If a subject using alphabetical organization can recall more words than an otherwise comparable subject who uses less efficient methods, we do not know whether this difference is attributable to differences in retention or only in recall. It is possible that both subjects have retained the same number of words and that the alphabetical organizer simply has access to a larger number of retained words, but it is also possible that the alphabetical organizer in fact has been able to store more information in his memory, that therefore more information is available in the storage, and that his superior recall merely reflects the greater amount of information that he has stored.

The logical requirements for the analytical separation of retention and recall, or the distinction between availability and accessibility of stored material, are simple enough. All that we need to do is to present the material to be remembered to different subjects under identical conditions, including identical instructions, and then vary experimentally the number of retrieval cues at the time of recall. If the groups are treated identically up to the beginning of the recall test, we can assume that the amount of information they have stored and thus the amount of information available is the same for all groups. If the number and kinds of retrieval cues given to different groups at the time of recall produce differences in recall, these differences would reflect differences in accessibility of the stored information and not differences in availability.

In the initial experiment we did to distinguish between availability and accessibility (Tulving and Pearlstone, 1966), we used a dual-level input list, consisting of category names and appropriate words as instances of these categories. Thus, for instance, one of the lists consisted of twenty-four words in twelve categories, with two words per category. The list presented to subjects might contain items like this: things found on a farm – WHEAT, TRACTOR; substances for flavoring food – SUGAR, CLOVES; geological formations – PLAIN, LAKE; musical instruments – DRUM, FLUTE; etc. Subjects were instructed that their task was to remember as many words as possible, that they would be tested for the words only, and that they did not have to memorize the category names. All subjects learning a particular list were treated identically up to the beginning of the recall test. At that point, half the subjects were asked to recall all the words they remembered, i.e, words other than the category names, while the other half were presented with the category names as retrieval cues and asked to write down all the words belonging to each category that they remembered from the list.

The design of the experiment included lists of three different lengths (twelve, twenty-four, and forty-eight words) and three levels of numbers of words per category (one, two, and four). Within each of these nine conditions of presentation, there were then two conditions of recall, cued and non-cued, as I have just described. The experiment was done with high-school students of both sexes as subjects, and data were collected in regular class sessions. Input material (category names and words to be remembered) was presented by means of tape recorder. The subjects wrote their recall in appropriately designed booklets. The presentation time varied with the length of the list and with the number of categories according to the formula: $T = NoC + L$, where T is presentation time in seconds, NoC is the number of categories, and L is list length. The amount of time given for recall was 1, 2, and 4 minutes for lists of twelve, twenty-four, and forty-eight words, respectively. A total of 949 subjects was tested, with approximately fifty-three subjects for each of the eighteen conditions of the experiment.

The mean number of words correctly recalled under each of the conditions of the experiment is shown in Table 1, together with corresponding standard deviations. As these data show, cued recall was superior to non-cued recall for all nine lists, the superiority varying directly with list length and inversely with the number of words per category. The difference between cued and non-cued recall thus was smallest for the list in which twelve words were organized into three categories, with four words per category. In this case, presentation of category names as retrieval cues apparently does not help subjects very much since they presumably can remember the category names on their own. The largest difference was found for the 48-word list in which each word belonged to a different category. Presentation of category names as retrieval cues in this list facilitated recall considerably: subjects in the cued recall group recalled over thirty-five words on the average, while the subjects in the non-cued group recalled less than sixteen words.

The major implication of the results of this experiment is that the subject's failure to recall a word he was instructed to recall does not necessarily mean that information about the membership of the word in the list is no more available in the memory storage. It is available, but not accessible under the typical unaided recall conditions. If the subject receives cues about the "location" of the relevant information in the storage, he will "find" many otherwise inaccessible words. Thus, intratrial "forgetting" (Tulving, 1964) of at least some of the items does not reflect the failure of the storage mechanism, but only the failure of the retrieval mechanism. How does the retrieval mechanism operate?

TABLE 1

MEAN NUMBER OF WORDS RECALLED AND THEIR STANDARD DEVIATIONS FOR THE EIGHTEEN EXPERIMENTAL CONDITIONS

		Number of Words Per Category					
		1		2		3	
List Length	Recall Conditions	M	S. D.	M	S. D.	M	S. D.
12	Cued	10. 70	1. 57	10. 94	1. 12	9. 98	1. 55
	Non-cued	7. 70	1. 61	8. 13	2. 03	9. 31	1. 91
24	Cued	21. 70	1. 97	19. 31	2. 38	15. 11	3. 17
	Non-cued	11. 18	2. 56	11. 82	2. 54	13. 38	3. 08
48	Cued	35. 35	6. 42	35. 76	7. 86	29. 60	6. 00
	Non-cued	15. 57	3. 53	18. 79	5. 99	19. 33	5. 34

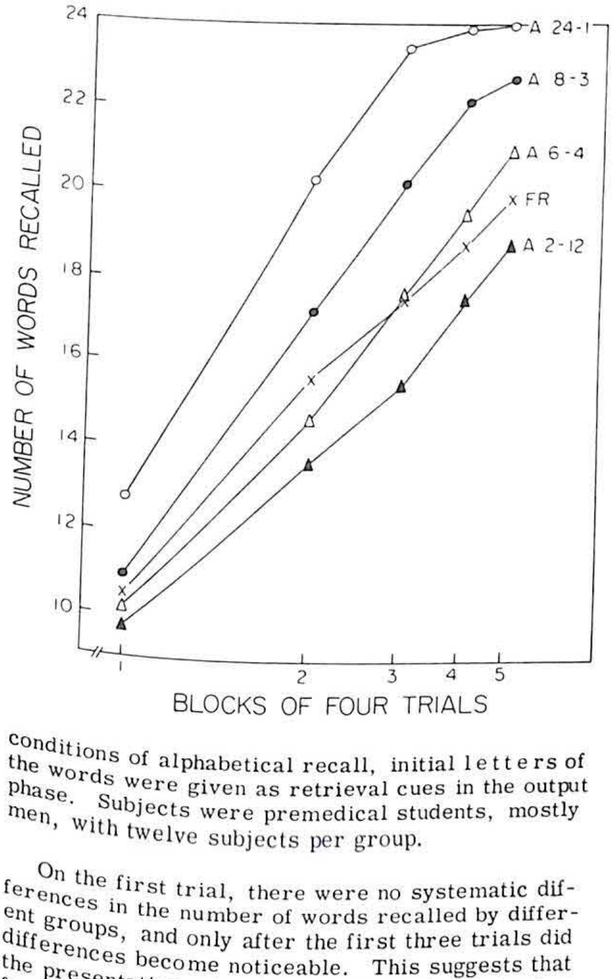
The Tulving and Pearlstone (1966) experiment provides some data relevant to this question. Word-recall data in an experiment of this type, where words are organized into higher-order units by the experimenter - in this case, into superordinate categories - and where this organization is adopted by the subject, can be analyzed into two independent multiplicative components, number of categories recalled and number of words recalled per recalled category (cf. Cohen, 1963, 1966). A category is said to be recalled - to be accessible - if at least one word from that category is recalled. Given the number of total words recalled and the number of categories recalled, the mean number of words recalled within recalled categories can be obtained simply by dividing the former number by the latter. For instance, if the subject recalls the words SUGAR, CLOVES, and LAKE, he has recalled two categories - *substances for flavoring food* and *geological formations* - and an average of 1. 5 words per category, two words from one category, and one from the other.

The breakdown of the word-recall data into these two components revealed two interesting facts. First, presentation of category names as retrieval cues affected only the number of categories recalled. In all lists the mean number of words recalled within a recalled category was the same under the cued and non-cued recall conditions. Second, word-recall increased as a direct function of list length for both cued and non-cued recall conditions, but again this increase was attributable primarily to the increase in the number of recalled categories and not to the increase in the number of words recalled per category. For the 24-word and 48-word lists the number of words recalled per category was invariant.

These data suggest then that recall of words from accessible categories is independent of the number of categories recalled, regardless of whether differences in category recall are brought about through presentation of category names as retrieval cues or through increasing the list length and thus the total number of categories presented. Retrieval mechanism seems to operate in some sense independently at two levels: at the level of higher-order units of organization and at the level of individual items within these higher-order units. Limitation of recall is determined by limitation of retrieval at both levels.

A further demonstration of the relation between number of categories and recall is provided by another experiment of Ozier's (1965). In this experiment subjects learned lists of twenty-four words either under conditions of free recall or under conditions of alphabetical organization. For different groups of subjects the number of unique initial letters of words in the list was varied. For one group, all words began with different letters. For another group there were twelve pairs of words (words within each pair beginning with the same letter); for a third group, eight initial letters were used, each associated with three different words, and so on, to the group in which twelve words began with one letter and twelve words with another. Thus, the experiment can be thought of as an analogy to the Tulving and Pearlstone experiment, with list length held constant at twenty-four words, the number of categories (initial letters of words) varying over the range of two to twenty-four, and the number of words per category (words sharing the initial letter) varying correspondingly over a range of twelve to one. Ozier's experiment differed from the Tulving and Pearlstone experiment in that it was a multitrial learning experiment. Each subject was given twenty trials of practice on the list to which he had been assigned. The order of words was randomized on all trials, the words were projected on a screen at the rate of one word/second, and subjects had 95 seconds for written recall on each trial. Under the

FIGURE 5. Mean recall on successive blocks of four trials for lists varying in number of different initial letters and learned under alphabetical organizing instructions (A), and for a list learned under free-recall instructions (FR). In the designation of A lists, the first figure refers to the number of different initial letters of words, and the second represents the number of words beginning with a common letter.



conditions of alphabetical recall, initial letters of the words were given as retrieval cues in the output phase. Subjects were premedical students, mostly men, with twelve subjects per group.

On the first trial, there were no systematic differences in the number of words recalled by different groups, and only after the first three trials did differences become noticeable. This suggests that the presentation rate of 1 second/word was too fast for words to be cued with respect to their initial letters. The learning curves for four alphabetical groups and a free-recall group are shown in Figure 5. The mean number of words recalled is plotted in Figure 5 against five blocks of four trials each on a logarithmic scale. These curves show that learning varied directly with the number of categories (number of unique initial letters) or inversely with the number of words per category. It is also interesting to note that free recall is more efficient than alphabetical recall of the list in which there are only two categories and twelve words per category.

The results of the last two experiments suggest that under conditions where the subject is provided with efficient retrieval cues – or where he can easily generate them from his own memory– and where the inter-item associative strength of items within a category is minimal, the optimum strategy is to organize the items into a maximum number of categories. On the other hand, under conditions where the subject can benefit from organization of individual items into higher-order units only to the extent that he himself remembers what these higher-order units are, the optimum strategy is to organize items into relatively few categories. It is most likely, however, that when the number of categories is decreased beyond a critical limit the difficulty of retrieval of items from within a given category will outweigh the facilitating effects of smaller number of categories, and recall will decrease. In the limiting case where all items in a list belong to a single category, recall is no better than recall of a list in which each item belongs to a different category.

SUMMARY

I have argued that understanding of many phenomena of verbal learning and memory would be enhanced if we adopted the basic orienting attitude that a person's mnemonic performance depends to a very large extent on the methods that he wittingly or unwittingly uses to organize the material. The important function of such organization is to facilitate retrieval of information from the memory storage.

I considered in some detail the problem of the effect of repetition on recall of a set of discrete items. Mere mechanical repetition of items does not seem to have any noticeable effect on the subsequent directed memorization of these items, provided that individual items as such are already well known to the subject before the experiment. Learning of the set, however, can be considerably speeded up if the subject uses effective methods of organizing the material. Individual differences among subjects in various learning tasks may reflect primarily their knowledge of, and the ability to utilize effective strategies and mediational devices, and only secondarily their innate differences in memorizing ability.

Different methods of organizing material may be incompatible with one another. When the subject has been using one method and then is asked to apply some other method to the material he is memorizing, the adjustment to the new method requires some time. Interference with overt recall under these conditions reflects incompatibility of different types of organization, and adds further support to the hypothesis that learning is critically dependent upon how the material to be remembered is organized.

It is not necessary to assume that organization changes the amount of material that the subject can retain, or the amount of information potentially available in the memory storage. Evidence suggests that organization only serves to make more of the available information accessible. The learner has always more mnemonic information available, relevant to the successful accomplishment of a memory task, than he can ordinarily make accessible to recall. Accessibility of information depends not only on its availability in the storage, but also on retrieval cues. Retrieval cues may be part of the environmental conditions prevailing at the time of recall, or they may have to be carried in memory. Limitation of the retrieval mechanism thus seems to be determined by the lack of accessibility of efficient retrieval cues, and the functional significance of organization lies in its role of providing such retrieval cues.

Optimum organization depends on the requirements of the recall task. If retrieval cues have to be generated from memory, the material is best organized into a relatively small number of categories – Mandler (1967) suggests that this number for common words is approximately five – if, on the other hand they will be available in the perceptual environment at the time of recall, the number of categories should be as large as possible. Retrieval cues such as category names provide access to categories or other kinds of higher-order units of material, but once access to such higher-order units becomes possible, retrieval of individual elements within units is unaffected by the number of higher-order units already recalled or yet to be recalled.

The evidence on which this account has been based is admittedly rather fragmentary, but the initial results are quite encouraging. Certain consistencies have already become apparent and further work will probably clarify those aspects of the conceptual framework that are still obscure.

While all this research that I have discussed was not undertaken with a view to solving practical problems as they occur in learning and remembering in real-life educational settings, it does have some implications for these problems. I will review just one, fully realizing that it is simply a reiteration of what many people have said before.

The kinds of findings from the simple experiments I have described in this paper once more underscore the need for the examination of methods that learners in classrooms use when they acquire knowledge and skills. We can never be certain that the methods of learning a person uses are the optimum ones for the requirements of any given task, or that these methods are sufficiently flexible to fit all tasks which in fact do require different strategies. And at the present time, we probably could not even evaluate the appropriateness of different approaches to different tasks or different subject matters, since we know so little about which methods are optimal for which kinds of learners in which situations. It seems to me that educators and psychologists have been more interested in methods of teaching than in methods of learning. Teachers are frequently taught how to teach, but how often are learners given specialized instruction in how to learn? A good teacher may be successful because, among other things, he organizes the material to be learned for the learner, but a good learner is likely to be more successful because he can organize the material on his own even if it is not optimally structured by the external environment.

To the extent that methods of teaching and methods of learning can be separated, it would seem that in the final analysis the latter are more important than the former, if for no other reason then at least because it might have greater transfer value as far as the individual learner is concerned. Even a poor learner, one who does not know how to use his innate intellectual abilities most effectively in learning situations, can do better if he has a good teacher to organize the material for him, but he cannot always expect to have good teachers around him. In many situations he has to be his own teacher. Would it be too rash to suggest that instruction in how to learn and how to remember be as important a part in the curriculum of the schools of tomorrow as are the three R's now?

FOOTNOTE

1. Research reported in this paper has been supported by Grant No. APT-39 from the National Research Council of Canada, and Grant No. GB-3710 from the National Science Foundation.

REFERENCES

Asch, S. E., and Ebenholtz, S. M. The process of free recall: Evidence for non-associative factors in acquisition and retention. Journal of Psychology, 1962, 54, 3-31.

Bousfield, W. A., Puff, C. R., and Cowan, T. M. The development of constancies in sequential organization during repeated free recall. Journal of Verbal Learning and Verbal Behavior, 1964, 3, 489-495.

Bugelski, B. R. Presentation time, total time, and mediation in paired-associate learning. Journal of Experimental Psychology, 1962, 63, 409-412.

Cohen, B. H. An investigation of recoding in free recall. Journal of Experimental Psychology, 1963, 65, 368-376.

Cohen, B. H. Some-or-none characteristics of coding. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 182-187.

Ebbinghaus, H. Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie. Leipzig, Drucker und Humblot, 1885.

Ehrlich, Stéphane. Le rôle de la structuration dans l'apprentissage verbal. Psychologie Francaise, 1965, 10, 119-146.

James, W. The principles of psychology. Vol. 1. New York: Holt, 1890; reprinted by Dover Publications, Inc., 1950.

Mandler, G. Subjects do think: A reply to Jung's comments. Psychological Review, 1965, 72, 323-326.

Mandler, G. Organization and memory. In K. W. Spence and Janet T. Spence (Eds.) The psychology of learning and motivation. New York: Academic Press, 1967.

Miller, G. A. Human memory and the storage of information. IRE Transactions on Information Theory, 1956, 2, 129-137. (a)

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97. (b)

Miller, G. A., Galanter, E., and Pribram, K. H. Plans and the structure of behavior. New York: Holt, 1960.

Miller, G. A., and McGill, W. J. A statistical description of verbal learning. Psychometrika, 1952, 17, 369-396.

Murdock, B. B., Jr. The immediate retention of unrelated words. Journal of Experimental Psychology, 1960, 60, 222-234.

Ozier, M. Alphabetic organization in memory. Unpublished doctoral dissertation, University of Toronto, 1965.

Pollack, I., Johnson, L. B., and Knaff, P. R. Running memory span. Journal of Experimental Psychology, 1959. 57, 137-146.

Thorndike, E. L., and Lorge, I. The teacher's word book of 30,000 words. New York: Bureau of Publications, Teachers College, Columbia University, 1944.

Tulving, E. Subjective organization in free recall of "unrelated" words. Psychological Review, 1962, 69, 344-354. (a)

Tulving, E. The effect of alphabetical subjective organization on memorizing unrelated words. Canadian Journal of Psychology, 1962, 16, 185-191. (b)

Tulving, E. Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. Psychological Review, 1964, 71, 219-237.

Tulving, E. Subjective organization and effects of repetition in multi-trial free-recall learning. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 193-197.

Tulving, E., and Pearlstone, Zena. Availability versus accessibility of information in memory for words. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 381-391.

Underwood, B. J., and Schulz, R. W. Meaningfulness and verbal learning. New York: Lippincott, 1960.

Waugh, N. C. Free versus serial recall. Journal of Experimental Psychology, 1961, 62, 496-502.

Developmental Processes in Thought

Kenneth Lovell
University of Leeds
Leeds, England

WHEN Professor Klausmeier chose the title for my paper, he asked me to deal with three points. First, give an outline of the conceptual framework that I am going to use to indicate developmental processes in thought. Second, give some account of the research that has been carried out to test the validity of the general framework. Third, point out the educational implications of what is dealt with in the first part of the paper.

The treatment of intellectual processes, including thought, is rather different in American psychology from that in either Russian or Swiss psychology. Only the future can show the extent to which these treatments can enrich one another, or the extent to which the different approaches say the same thing using different terminologies. Individuals wishing to know more of these three approaches would, no doubt, be interested in Berlyne's most recent book *Structure and Direction in Thinking* (1965). However, I shall deal with my topic largely from the standpoint of the Genevan school, for Piaget has great influence in British educational circles, although he is not quite so influential among British psychologists.

FIRST 21 MONTHS OF CHILD'S LIFE

There is now much data which indicates that the first 21 months or so of a child's life are of great importance from the point of view of the growth of thought. Piaget has shown that a baby's movements are not random; indeed, on close inspection they are seen to conform to a pattern. This period of development was called by Piaget the period of *sensori-motor* intelligence, because the schemas, that is the sequence and the structure of the actions slowly built up in the mind, are dependent upon the direct support of perception and motor activity. From 12 to 18 months of age the child goes through a period of intense experimentation. He will let an object fall to the floor to see what happens; he will knock things, shake them, or throw them to see if they roll. New ways of doing things are discovered based on a grasp of new relationships. For example, a child will pull a box toward him in order to obtain a toy placed on it, which otherwise would be out of reach. And between 18 and 21 months of age, the awareness of relationships is sufficiently developed for the child to be able to invent new means, and to be able to foresee which acts will succeed and which will not, without putting them to the test. This invention comes through a covert process that amounts to internal experimentation. The earlier trial and error period is not now so important, for the child can increasingly represent to himself the various possible actions and how they must be combined, to attain a desired end. The child is beginning to think, for his actions are now carried out in an inward form. Some psychologists would call such internalized actions, incipient or implicit responses.

During the first 21 months or so of life, the ordinary child elaborates the basic schemas of the object, of space, of time, and of causality. For example, the basic schemas in relation to space and time are laid down as when the child adjusts his reaching actions for near and distant objects, and when he moves to catch a swinging rattle.

In Piagets's view, the child in adapting himself to his environment has to assimilate or absorb new experiences into his existing schemas. At the same time he must accommodate to the environment through the modification of existing schemas or the build up of new ones. Cognitive adaptation to some new situation, and hence the ability to deal with it in thought, involves both these processes. To assimilate the meaning of a new situation the child must accommodate to it; and in order to accommodate to it, he must be able to assimilate it. Cognitive growth is thus a step by step process with the new always building on the old. But once new experience is assimilated, the child's schemas become more complex, and because of this, more complex accommodations are possible. Moreover, the child's schemas do not remain unchanged even in the absence of environmental stimulation, for meanings are constantly reorganized, and linked with other meanings. This internal renovation is, in Piaget's view, an important source of cognitive development. Even so, intellectual growth is slow. We must also note that Piaget's observations lead him to believe that it is schemas in the process of organization that children tend to repeat playfully and with seeming pleasure. But when such schemas have become organized, the apparent pleasure disappears and they cease to be repeated unless they are combined to form new schemas or serve as a means to some end. Flavell (1963) remarks that schemas are structures, and one of their important, built in properties, is that of repeated assimilation of anything assimilable in the environment.

BETWEEN 2 AND 8 YEARS OF AGE

Throughout the sensori-motor period the child is unable to use an image or word, which could represent to him an object or event not actually present. For Piaget, the child moves out of this stage into what he calls the pre-operational stage of thought when the child can differentiate a word from what it stands for; for example, use the sound "dog" to stand for, or represent, a dog. When he can do this he can represent to himself situations that are no longer in actual evidence, and thought is lifted to an entirely new level. Note carefully, however, that thought comes before language; the latter is fitted on to thought that already exists. But once the child can use language, thought is extended over an immensely increased range. Moreover, he now has a more permanent and far more flexible model of the outside world, for he is no longer dependent upon immediate perception and motor activity for thought. Nevertheless, between 2 and 4 or 5 years of age, thinking tends to center on one striking feature of the situation; it is also irreversible in that the child is unable to move back, in his mind, to the starting point from which his immediate thinking began. The child at this stage can only see an event from his own viewpoint; he is unable to conceive the viewpoint of other children. Yet by 3 or 4 years of age the child has copied the adult model of language, and at this age we are in grave danger of overestimating his level of thought from the nature of his speech.

By about 5 years of age a change begins to set in, and by 7 to 8 years of age the child's thinking becomes more systematized; that is, his thoughts now conform to certain rules and his thinking becomes what we adults call logical. The sequence and structure of actions in the mind, or the schemas, now available are now altogether different in kind. The ability to reason and "understand" demands higher order schemas which permit a simultaneous grasp of the successive sequences of actions taking place in the mind. The ordinary child of 7 to 8 years can, as it were, "look down on his schemas" or "turn round on his schemas." He is then aware of the sequences of action in his mind; he can see the part played by himself in ordering his experiences; and for any action in his mind, he can see that there are other actions that give the same result. That is, he sees equivalences. Thus he understands that $4+2=3+3=7-1$ or that all girls plus all boys plus all adults = all boys plus all adults plus all girls. Indeed, there is now a new kind of coordination of schemas, yielding a simulataneous understanding of equivalences for actions within the mind. Thus the child can now measure the same distance in feet and inches and understand that the different figures mean the same thing; he can perform subtraction using the method of complementary addition; he can decompose 42 units in 4 tens and 2 units with understanding. Thinking now conforms to a system, and there is learning with a certain amount of understanding.

At the level of sensori-motor and pre-operational thought there was a great deal of learning, but little or no understanding in the sense that the child learned a linear sequence of actions but was unable to elaborate a set of equivalences. We must not in any way belittle learning that takes place before thought becomes systematic, for such learning is essential for the growth of thought itself. Indeed in older normal children, and in adults, a great deal of learning is of this type.

Now because the elementary-school child can increasingly see the part played by himself in ordering his experiences, and because he can increasingly coordinate actions in his mind and make them conform to a system, he can build the concepts of a class, a series, length, time, and so on. But the concepts that he builds are only those that he can attain from experience with first hand reality, and his systematized or logical thought is related only to the world of perceivable things and events. For Piaget, the elementary-school child is at the stage of concrete operational thought and is, for example, able to understand certain aspects of mathematics.

END OF ELEMENTARY SCHOOL PERIOD

Toward the end of the elementary-school period, however, children begin to realize that there are gaps and uncertainties in their thinking, and that certain kinds of problems cannot be solved by them. As the child becomes better at organizing and structuring problem data with concrete operational methods, he becomes more aware that the latter does not yield a logically exhaustive solution to his problem. He gropes for new methods of attack, often in Piaget's view, as the adolescent commits himself to real life situations.

So from 11 to 12 years of age in able children, and from 13 to 14 in ordinary pupils, new thinking skills begin to emerge. Due to his continued interactions with the cultural milieu, and to the maturation of the central nervous system, the individual can produce more complex expectations when faced with certain kinds of situation and data. His logical thought is no longer restricted as before, for he can increasingly use statements or propositions relevant to objects and their relations. The sequence and structure of his mental actions, or his schemas, are new in kind. The pupil can now begin to manipulate statements which refer to classes and relations. For example, he can answer the question "If Henry is taller than Mary and shorter than Debbie, who is the tallest?" Or he can argue from the situation in front of him that, "Either this must be true or that must be true," or he can see that "X implies Y." The pupil can, for any sequence of actions in his mind, see a greater range of equivalent actions than was the case at elementary-school level. Faced with certain situations the child can, at 13 or 14, set up an hypothesis and work out what would happen if this hypothesis was true. He is now able to deal with the merely possible. The pupil has reached what Piaget calls the stage of formal operational thought, and many experiments to illustrate the growth of such thought are to be found in Inhelder and Piaget's (1958) book The Growth of Logical Thinking from Childhood to Adolescence (1958). A number of these experiments were repeated by Lovell (1961) and Jackson (1965), and the broad stages in the growth of logical thinking were mainly confirmed although the position is more complicated than Inhelder and Piaget suggested. It must be stressed, however, that formal thought is found in all areas of thinking – in history, literature, politics – and not just in science and technology. There is much published work to support this statement, and if space permitted, I could give you some interesting illustrations for formal thinking in these fields.

With the onset of formal thought, the pupil is able to elaborate an entirely new kind of concept. You will remember that in the elementary school the child could say how he ordered his experience, for he could dissociate the part played by himself in classifying his experience from the characteristics of that experience. But in adolescence the pupil is able to structure and coordinate actions upon relations, which themselves result from the coordination of actions. For example, in the case of heat in physics, the concept depends on the earlier elaboration of the concepts of mass and temperature. The latter are developed at the level of concrete operational thought, for each is a coordination of some intuitive aspect of reality, but their product is not. Concepts derived at the level of formal thought depend upon the concepts elaborated at the level of concrete thought being completely detached from their concrete contexts and manipulated as "pure" concepts.

BEHAVIOR THEORY

Many individuals have been brought up on what may be broadly called behavior theory, although not all Americans think of cognitive growth in behavioral terms. Some are now talking of "strategies"; "structure" and "rule"; "programs"; "plans." But for people who have studied behavior theory I would like to review a few points that will help them grasp how the simplest principles of behaviorism look to a Piagetian. First, if two schemas are frequently evoked in regular succession they tend to form a single, larger, schema comprising the first and then the second in that order. The new and larger schema is then evoked as a whole. This is "conditioning." The individual action is, of course, the limiting case of a schema. Second, if two or more schemas are frequently evoked in irregular succession they, too, tend to form a larger schema made up of all the original schemas. This larger schema will also be evoked as a whole, although the order of the evocation of its parts will depend upon the stimulus properties of the situation. Nevertheless, its value, as a larger schema, lies in the mutual inter-facilitation of its parts. This provides the basis for "trial and error" behavior thought by Thorndike to be the prototype of all learning. Third, if a situation evokes a larger "trial and error" system of actions or schema as indicated in my second point, and the larger schema includes a particular sub-schema that is particularly appropriate to the situation, the effect of accommodation will be to inhibit and finally eliminate the other sub-schemas and so strengthen the particular sub-schema (Thorndike's Law of Effect).

The functional principles indicated under my three points seem to be common to all schemas and show the most elementary forms of coordination of schemas. They may account for habit formation, or for simple kinds of learning, either in humans or in animals, which involve a linear sequence of actions. But we do not know if they could be made to account for understanding, for the capacity to reason and to understand are not functions or properties of all schemas. From time to time a combination of my first two points may well give more "intelligent" trial and error behavior from which outsiders or onlookers may infer that two separate sequences of actions are equivalent to one another. For example, a young child, or an older retarded child runs into the house by the shortest route which is the one normally used. But if he finds the gate locked, he at once selects the best alternative. Equivalence is implied by the detour behavior, but it is not explicit in the child's representation of the situation.

Within these first and second order operational schemas, i.e., schemas that function at the level of concrete and formal operational thought respectively – the operational structure of systems of equivalences form a sub-schema of great generality and enter into all processes of learning. The acquisition of such schemas are essential to learning. Such schemas do not arise merely by the pupil knowing if his answers are right or wrong; rather they seem to depend upon an active coordination by the subject of his existing schemas – on his ability to order these actions simultaneously and to recognize their systematic equivalences. Good teaching presents situations in which such coordination is possible and necessary, and in enough variety for the pupil to dissociate what is common, although it cannot ensure this. It will readily be realized that not all learning either in elementary or high school is of this form. When a 9- or 15-year-old is presented with a problem for which he has an adequate schema, he will assimilate the problem to the schema by an appropriate method of attack and arrive at a solution. This is learning of content. He has not acquired a new method of coordination for he already had the method. All the problem did was to strengthen the tendency for this schema to be evoked when faced with problems of this type.

The most elementary ways in which schemas can be coordinated have just been indicated and all schemas probably lend themselves to this form of coordination. But since these simple methods of coordination do not involve a simultaneous apprehension of successive schemas by higher order schemas, the connections implied by the coordinations are linear. We thus have a linear series of actions at the first level of abstraction involving "learning" but little understanding. Most learning in the mentally retarded, and much in ordinary humans, is of this type.

Let me emphasize that I am well aware that I have not allowed myself to get bogged down over the question of whether or not the stages of cognitive growth that I have indicated do actually exist in the child's intellectual growth. Piaget has been much criticized over his stages, some having argued that stages do not exist in other aspects of growth. The current practice among researchers of subjectively deciding whether responses can be fitted into Piaget's schema of stages is obviously not suitable for assessing the validity of these stages. More objective methods are required which would allow the data to determine their own patterns and groupings and to confirm or deny the existence of stages.

FINDINGS FROM RESEARCH ON PIAGET'S IDEAS

The second part of my paper is about the findings from the research that has been carried out to test Piaget's ideas. I have heavily condensed the literature above since the material now at hand is very great. Studies have been made in the realm of the general growth of logical thought, into the child's ability to classify and serialize, and into his notions of space, time, causality, and the like. Whatever the future holds regarding the presence of stages, it can be said at the present that research shows that a high proportion of protocols obtained from pupils can be put into the stages proposed by Piaget, or into intermediate stages. It is true to say that the broad picture of the growth of thought, presented by Piaget, has been confirmed, although there are many blots on the canvas, and at times one must be extremely critical of his work. But all in all he has done more than anyone else to throw light on the growth of thought processes. His experiments will stand the test of time, although his theoretical system is certain to undergo changes.

The great number of studies relating to the growth of concepts shows that it is between 6 and 8 years of age that the pupil begins to elaborate the basic concepts of mathematics and science for he can now dissociate the part played by himself in ordering his experience. If we take, say, the concept of time, it has been amply confirmed that the child begins to coordinate instants and intervals around 8 years of age, although he may well have been using time words and telling the time for some years previously. Likewise experiments relating to the overall growth of logical thought have shown the same general progression whether in the area of science, history or religious thinking. The available data again confirms that it is not until 13 to 14 in ordinary pupils and 11 to 12 in the very able, that formal thought is possible and second order concepts are elaborated. I must stress, however, that the stages are not clear cut, and I shall return to this point. Indeed the stages are blurred, so that when a pupil is approaching the stage of concrete or formal thought, he is found to be at a more advanced stage of thinking in one situation than he is in another although both involve the same operational structure. It is also true that in some experiments certain stages appear to be absent; some responses are found that cannot be fitted into Piaget's stages; and occasionally a child is found who appears to conserve, say, weight before substance. On the other hand, there is no research that I know of, that has shown that Piaget was wrong concerning the sequence of stages for any one situation for any one child. This was clearly seen in a recent study, just completed, where a group of boys was tested on the same experiments each year between 11 and 15 years of age.

It would not be inappropriate if I now gave you a couple of examples of protocols obtained in a recent study in the growth of logical thought in relation to history. It will keep our discussion well and truly about children. I take the example from history as I cannot then be accused of being interested only in mathematics and science. Pupils read a story concerning William the Conqueror – the last man who successfully invaded England and that was in 1066. With the story still in front of him, each child was asked individually, "Was William the Conqueror a cruel man?" This was followed by "Why do you think so?"

One pupil age 13 years 8 months gave a reply that was clearly at the level of concrete thought. He said, "Well, William wasn't cruel at the beginning because he allowed the English to collect the dead bodies. If he was really a cruel man he wouldn't even have allowed that. Yet later on, after the Danes had left, he'd no need to take revenge. He could have reconquered the North – tried to make treaties and things." Here the pupil could use the information provided, but he was not able to form a mature hypothesis from a consideration of all the implications in a situation.

Another pupil age 14 years and 8 months replied at the level of formal operational thought. "It depends on what you call cruel. If the definition of 'cruel' is to kill and ravish and burn for any purpose whatever, William was cruel. On the other hand, if one is prepared to accept political necessity, William's cruelty was justified. Compared with many other feudatories, knights and so on, he was essentially a kind man. They ravaged generally for their own advantage and without care for the common folk of the land. Duke William, if the common people went with him, seems to have been prepared to protect the common people from ravages. If, however, they went against him, he seems to have treated it as a deliberate breaking of faith and acted accordingly. So, by the standards of his own day – for we really cannot judge him by our standards –

he was probably not a 'cruel man.' " This pupil was clearly able to reason by implication of an abstract lead.

One of the consistent findings in our studies at Leeds, and it has been supported by the work of Jackson (1965) – a student of Dr. Lunzer of Manchester – is that only a proportion of school educable retarded pupils reaches the level of concrete operational thought even at the age of 15. It has repeatedly been shown to be true in respect of mathematics and science; while recently we have shown that the growth of English morphology in such children, even at 15, is less advanced than that of normal children of 4 and 5 years of age if we accept Berko's (1958) data as representative of American children. Our findings, which will be published, were predicted from our knowledge of the thinking of this type of child in other areas. The learning problems of school educable retarded children, and of the least able left in elementary and high school, stem from the fact that their schemas permit only a linear series of actions below, or just at, the level of concrete operational thought. They learn, but they do not "understand." Transfer certainly takes place in these pupils as it did in Harlow's apes, but only when the schemas involved are less complex than those required at the level of concrete operational thought.

One great issue that research has clearly indicated is that problems that appear to have the same operational structure are not all solvable at the same time. Much research shows that both concrete and formal operational thought is, at first, a task specific within limits, and that thinking is not organized to the extent that Piaget's theory would indicate. Dodwell (1960, 1961) showed that there was only a moderate correlation between tests which, in Piaget's view, all involved the same level of thinking and which all involved integral aspects of the number concepts. Later Dodwell (1962) pointed out that while the concepts of class and number develop within the same age range, there was no clear indication that they both arose together or that one was elaborated before the other. Likewise we at Leeds have found, as did Dodwell (1963), that a pupil could be at different stages of thought in three tests dealing with the concept of axes of reference. Indeed, Piaget (1960) himself has now admitted that operations are only gradually applied to larger and larger numerical sets, while he has always made it clear that it takes a child, on the average, two years to generalize concrete operations involved in appreciating conservation of quantity before they can be applied to conservation of weight, although from the point of view of cognitive structure the operations are the same. Even when the concept involved remains precisely the same, the quality of the child's thinking will vary according to the apparatus used (cf. Lovell and Slater, 1960).

It would, of course, be unfair to the Genevan workers if one failed to emphasize that they have made some provision for these eventualities. Inhelder and Piaget (1958) pointed out that concrete operations consist of the direct organization of immediately given data and they cannot be generalized to all situations at once. For example, length is conceived before weight. This, in the view of Inhelder and Piaget, is because it is more difficult to serialize, equalize, etc., objects whose properties are less easy to dissociate from one's actions, e.g., weight, than to apply concrete operations to properties that can be rendered more objective, e.g., length. Piaget (1956) also speaks of the notion of "horizontal differentials." This suggests that the same or similar concepts when derived from different materials or situations, develop in staggered sequence rather than simultaneously. But this notion does not fit well into his general theory.

At the level of formal thought we find much the same. In a study recently completed, in which children were questioned individually, we have found that the schema of proportion is not available in problems involving money, speeds, areas, series, etc., at the same time. Mathematical concepts seem often to be available at first in specific situations, even when they depend on second order operational schemas. Lack of specific experience, information, vocabulary, expectancy or individual differences in intellectual functioning which are unknown, most probably all play some part.

A number of studies have been undertaken, based on a neo-behaviorist approach, to see if intensive periods of specific training lasting a few weeks, can speed up the growth of understanding of particular issues; for example, the conservation of number or weight. The outcome of these studies has been largely inconclusive. Piaget's view (Ripple and Rockcastle, 1964), which was expressed in 1964, is that the child may learn something of the situation, but the training will have no effect on his general level of understanding for the specific attack is too trivial. The modification of a child's mental structures in Paiget's view, necessitates a far wider, more lasting, and more radical approach which involves many of its child's activities.

THE WORK OF SMEDSLUND

In concluding this section of my paper I will review a few details of a recently published work of Smedslund (1964). It was a very important study as it is extrememly relevant to some of the points that I have raised in this section of my paper. The specific purpose of the study was to investigate the interrelations of the specific acquisition of ability for concrete reasoning using different items. He carefully laid down a number of methodological rules which were applied to the construction of items in order to maximize their diagnostic validity. The items included class inclusion, reversal of spatial order, conservation of discontinuous quantities, conservation of length, transitivity of length, etc., – nine items in all. Many of the items have sub-items. All the tests were given individually to 160 children age 4 years 3 months to 11 years 4 months, evenly distributed over age and sex.

The upshot of his findings was that under the given conditions of the experiment, one could predict with a fairly high degree of confidence from the success or failure on one sub-item to success or failure on another sub-item; that is, intra-item reliability was high. On the other hand, the picture was different between items, for the inter-item relationships were far lower. The difficulty of the items varied from 122 passes to 77 passes. When children who had passed or failed all the items were excluded, only 18 percent of the remaining subjects conformed with the hypothesis that the items are acquired in the order of their difficulty as measured by the total number of subjects passed.

Smedslund points out, in retrospect, what he considers to be a marked weakness in his study. He proposes that when investigating a concrete reasoning task we need to make a clear distinction between percept – a process depending upon the momentary stimulus input; goal object – that which the subject is told to attain; and inference pattern which is formed by the set of premises and the conclusion. For example, the stimulus situation as apprehended by the subject are the percepts; quantity and length are goal objects; and conservation and transitivity are inference patterns. Smedslund thinks that since all three factors influence the solution rate of items, the effect of a single factor can only be studied with the others held constant. The author is of the opinion that other things being equal, conservation precedes transitivity, but he feels that only in relation to one goal object, and one narrow perceptual context, will it be possible to determine whether or not conservation, transitivity, seriation, associativity, commentativity, addition/subtraction, multiplication, etc., are acquired in rigid predictable sequences. Here, indeed, is a field for research.

IMPLICATIONS

In this section of my paper, I want to discuss my interpretation of the educational implications of Piaget's work and to suggest onr or two lines of research. To be helpful I have listed a number of points.

1. While experience is always necessary for mental growth, Piaget is clear that mere submission to external experience is not sufficient to make a child restructure his own thinking. Rather the child has to be active. He has to act on material things and become aware of the significance of his actions.

2. His observations led him to believe that it is schemas that are in the process of organization that children tend to repeat playfully and with seeming pleasure. Further, when such schemas have become organized, the apparent pleasure disappears and the schemas cease to be repeated unless they are combined to form new schemas or serve as a means to some end. Learning then seems to start from the child, from the schemas that he already has available. Actions on his part, that is, exploring, discovering, using new ways to solve old problems, all have an intrinsic interest for the young, and are self-extending. Naturally the level of intrinsic motivation varies from child to child, due, perhaps, to the level of activity in certain areas of the central nervous system. But when the schemas required for the solution to some problem are not too far removed in complexity from those available to the child, the inadequacy of existing schemas will force him to accommodate to the conditions of the problem. Hence the child restructures his own schemas toward greater cognitive adaptation to his environment. Not only does the child solve the problem, but he extends his capacity for further learning. The implication of Piaget's findings is that we need to begin to stimulate the intellectual growth of the child in the early weeks of life. The child needs to encounter a variety of changes of sensory experience in respect of sight, sound, touch, and movement. He needs an opportunity to see and grasp a variety of objects, and later, suitable toys that involve him in a greater range of actions; opportunities for physical movement and experimentation; and novelty. We must always bear in mind that there must not be too great a gap between the schemas available to the child and those demanded by the situation. Yet in spite of the help given by Piaget in assessing a child's level of thinking, what is involved in producing the correct amount of gap between the schemas available to the child and those demanded by the situation remains vague. This is where the intuitive skill of the teacher is called for. It is his task to arrange, or find in the environment, problems which call forth the schemas of the child in new and novel ways.

3. While language acts as a vehicle for thought and enables it to be carried immensely further, language is fitted into the thinking that originates in the first two years of life. If thought and language are to keep in step with one another as the child continuously models his language on that of adults, the child must be active in building his thought processes. If this is not the case, it is more likely that the child's verbalizations will lead adults to think that he understands more than he actually does. There is a grave danger of this happening between about 2 and 5 years of age since the child has copied the adult model of language yet he remains at the pre-operational level of thought. One fundamantal lesson for parent and teacher from Piaget is that while language is important, mere verbalization and verbal knowledge are of little value in themselves.

4. Piaget (1926) indicated that social influences play a role in helping a child to move from cognitive egocentrism to systematic thought. Mere experience may have no effect on the available schemas, or an attempt may be made to distort the experience so that it could be assimilated. But social interaction with the peer group forces the child to re-examine his own thinking, so that he can satisfy his need to share the thoughts of others and his need to communicate with them. Argument with the peer group forces the child to reason with himself. Likewise, dialogue and discussion with parents and teachers, the answering of questions, etc., contribute to the growth of concrete operational thought.

5. Piaget's views clearly support what is broadly called a discovery approach to learning with a judicious use of exposition at the elementary school level. But in high school more use can be made of exposition since the child's ability to appreciate the form of an argument improves, and he is able to elaborate concepts not themselves derivable from first hand experience. The Geneva school argues that the onset of the stage of formal thought is relative to the culture pattern. Beyond the age of 12 years or so, due to physiological factors, this level of thinking may be a product of the progressive acceleration of individual development under the influence of education and culture. The 13 year-old is not only dissatisfied with the gaps and uncertainties resulting from concrete operational thought,

but he is beginning to think beyond the present. This thinking ahead is likely to be determined by the experiences received at home, at school, and at work; his social attitudes; the climate of opinion and expectancy in the community, and the concepts that are frequently made use of in the society. In committing himself to the future, the adolescent begins to build theories.

All high school pupils thus need the opportunities to discuss with adults and teachers, and among themselves, viewpoints and theories relating to varied problems; the viewpoints and theories sometimes being in direct conflict with one another. They also need the chance to see, at first hand, the kinds of employment that will be available to them, and be in a position to think of, and discuss, their future roles as workers and citizens. They thereby get greater opportunities to commit themselves to possibilities. The greater the need to question and find out, to struggle for solutions to problems, and to commit oneself to possibilities, the greater seems the likelihood of formal thought developing. It may be that the culture pattern rather than the school plays the greater role in the growth of formal thought although we do not know if this is so in fact. Peluffo's (1964) study, which compared children born in Sardinia but who had been in Genoa for varying lengths of time, with children born and bred in Genoa, gives some indication of the likely effects of the culture pattern on the onset of formal thought.

6. In the upper-elementary classes and in high school, extrinsic motivation begins to influence the pupil. For example, he may become aware of what his parents or society expect of him; or he may fear failure or punishment if he does not work hard. One must not belittle this kind of motivation for it often stimulates the child to learn and to work as hard as he can at the level of thinking at which he is. But there is no evidence that it can, in itself, change the quality of the pupil's thought. Extrinsic motivation may well hold a pupil to a task, and in virtue of the prolonged child-task inter-action enable him to restructure his own schemas. Thus external influences which are at work over a number of years, as, for example, parental encouragement, can play an important but indirect part in intellectual growth.

CONCLUSION

In conclusion, I would like to suggest areas in which we need research:

1. What are the long-term effects of very early stimulation of the culturally deprived and of certain types of school educable retarded children? If such is commenced during Piaget's sensori-motor stage of development, how do the long-term outcomes compare with those outcomes when the enriched environment is delayed until 3 and 5 years of age? The recently reported long-term study of Skeels (1965) in America suggests that very early stimulation may be of great importance. And what is the most suitable stimulation to provide at 6, 9, 15 months?

2. What are the long-term effects of an education based, as it were, on Piagetian ideas, particularly when they employ materials and situations of intrinsic interest to the child, from K through grade 4? Of course there will be some exposition, as there will be an increased proportion of time spent on exposition in the upper classes of the elementary school and in high school. What are the long-term effects of such methods over, say, 15 to 18 years? Is there any difference between such children and those brought up on other approaches at the second, fourth, eighth, and twelfth grades?

3. What is the effect of the emotional life on cognitive development? This is an important, although complex, question that cannot be brushed aside. In the 3 to 5 year period especially, fantasy - which I define as emotion clothed with images - is playing a role in the growth of thought. We ignore this at our peril. If in real life a child can dramatize his fantasy, he may free himself from the dominance of fantasy, and thought processes may develop smoothly. When the 2- to 5-year-old is presented with a story or situation which he only partly understands, he brings to it his experiences and his fantasies to fill in gaps and structure, say, the story. Each young child will have something of his own in the interpretation of the story, for information cannot yet be classified systematically and objectively. In the fairy tale particularly - which again is only half understood - the child's feelings of fear, anger, love, etc., which already exist, are perceived by him in the story. The story acts as a mirror and he is helped in recognizing the parts in himself in the "Good Prince" or "Bad Dragon." The "mirror-like properties" of fairy stories, and the fact that he can structure it with his own experience and fantasy, perhaps explains its abiding value. Similarly the child deals with his fantasy in painting or modeling. At 6 to 7 years of age fantasy is still at work, but the child is in better control. He may now accept a story that he could not tolerate at 4. I am no clinician in the accepted sense of the term, but I have enough experience with young children to know that the fantasy life is playing a role in the growth of thought processes although we have little evidence in a scientific sense on this point. What is the effect of kindergarten and school activities and which are likely to help the child to control his fantasy? What is the effect of using materials likely to have personal significance for him? The mathematician-logician may be forgiven for ignoring this when he is trying to teach kindergarten children something about sets. The psychologist and educator should realize that here is an important field about which we know little.

4. What is the effect of the culture pattern and subculture patterns on the growth of human thought processes? We know something of the relation between socioeconomic groupings and attainment but far less about such groupings and thought. Evidence from Hong Kong via Goodnow (1962), and from the Tiv tribe of Nigeria via Price-Williams (1961, 1962) suggests that culture pattern may not have as serious effect on the growth of concrete operational thought as on formal thought, especially if the materials used are well known to the testees. Nevertheless there is some delay at the level of concrete reasoning due to culture pattern as the work of Peluffo (1964) in Italy, and Almy (1966) in America. What are the characteristics of the sub-cultures that have the greatest downpulling effect?

5. We need to know why a schema, say, that of proportion cannot be applied to a wide variety of situations at the same time. How important is familiarity with the problem, lack of specific experience in that area, and expectations based on previous experience?

6. We need a very great deal of information regarding the growth of more advanced concepts in mathematics and science, e. g., "function" and "entropy."

These questions that I have raised are broad ones, but ones of great importance. They are of significance to mankind, not to just the Americans or the British.

REFERENCES

Almy, M. Young children's thinking. New York: Teachers College Press, Columbia University, 1966.

Berko, J. The child's learning of English morphology. Word, 1958, 14, 150-177.

Berlyne, D. E. Structure and direction in thinking. London, Wiley, 1965.

Dodwell, P. C. Children's understanding of number and related concepts. Canadian Journal of Psychology, 1960, 14, 191-205.

Dodwell, P. C. Children's understanding of number concepts: Characteristics of an individual and of a group test. Canadian Journal of Psychology, 1961, 15, 29-36.

Dodwell, P. C. Relations between the understanding of the logic of classes and of cardinal number in children. Canadian Journal of Psychology, 1962, 16, 152-160.

Dodwell, P. C. Children's understanding of spatial concepts. Canadian Journal of Psychology, 1963, 17, 141-161.

Flavell, J. H. The developmental psychology of Jean Piaget. London: Van Nostrand, 1963.

Goodnow, Jacqueline J. A test of milieu effects with some of Piaget's tasks. Psychological Monographs, 1962, 76 (36, Whole No. 555).

Inhelder, B., and Piaget, J. The growth of logical thinking from childhood to adolescence. London: Routledge, 1958.

Jackson, S. The growth of logical thinking in normal and subnormal children. British Journal of Educational Psychology, 1965, 35, 255-258.

Lovell, K. A follow-up study of Inhelder and Piaget's The Growth of Logical Thinking. British Journal of Psychology, 1961, 52, 143-153.

Lovell, K., and Slater, A. The growth of the concept of time: A comparative study. Journal of Child Psychology and Psychiatry, 1960, 1, 179-190.

Peluffo, N. La nozione di conservazione del volume e le operazioni di combinazione come indice di sviluppo del pensiero operatorio in soggett appartenenti ad ambienti fisici e socioculturali diversi. Rivista di Psicologia Sociale, 1964, 31, 99-132.

Piaget, J. The language and thought of the child. London: Kegan Paul, French, Trubner, 1926.

Piaget, J. Les stades du développment intellectual de l'enfant et de l'adolescent. In P. Osterrieth et al. (Eds.), Le problème des stades en psychologie de l'enfant. Paris: Presses Universitaires de France, 1956. Pp. 33-113.

Piaget, J. Introduction. In P. Gréco, J. B. Grize, S. Papert, and J. Piaget, Problèmes de la construction du nombre. Études d'épistemologie génétique. Vol. 11. Paris: Presses Universitaires de France, 1960. Pp. 1-3.

Price-Williams, D. R. A study concerning concepts of conservation of quantities among primitive children. Acta Psychologica, 1961, 18, 297-305.

Price-Williams, D. R. Abstract and concrete modes of classification in a primitive society. British Journal of Educational Psychology, 1962, 32, 50-61.

Ripple, R. E., and Rockcastle, V. N. (Eds.) Paiget rediscovered. Ithaca: School of Education, Cornell University, 1964.

Skeels, H. M. Effects of adoption on children from institutions. Children, 1965, 12, 33-34.

Smedslund, J. Concrete reasoning: A study of intellectual development. Monographs of the Society for Research in Child Development, 1964, 29, (2, Whole No. 93).

Promoting Creative Thinking in the Classroom[1]

Martin V. Covington
University of California
Berkeley

THE LAST decade has seen a sharp increase in the volume of research on creativity and creative thinking with a major focus on isolating the basic processes of creative thought (Mednick, 1962; Mendelsohn and Griswold, 1964) and on identifying the unique characteristics of creative persons (MacKinnon, 1965; Barron, 1964; Crutchfield, 1963). Despite such unprecedented interest in creativity, until quite recently there has been relatively little research on the general problem of nurturing and promoting creative thinking, especially in the classroom setting. However, due partly to increased public awareness of the need to husband our intellectual resources, this important topic is beginning to receive the attention and research support it deserves.

The purpose of this paper is to describe part of the research program at Berkeley for developing curriculum programs aimed at promoting the general level of creative thinking among elementary-school children and to consider some of the pedagogical implications of such research, especially regarding the value, feasibility, and justification of teaching for creative functioning. For a survey of other current research in this area representing differing theoretical and methodological approaches, see Myers and Torrance (1964), Osborn (1953), Parnes (1965), Suchman (1961), and Upton and Sampson (1961).

By way of introduction, it should be noted that most of the current attempts to develop such teaching materials, including that of the Berkeley group, share two fundamental assumptions: (1) that all children, regardless of age or initial intellectual level, fall far short of realizing their potential for creative thought, and (2) that these individuals can be taught to make fuller and more appropriate use of their capacities by means of systematic and direct training of certain cognitive skills.

As for a definition of a creative child (in a real sense the ultimate criterion by which the success of a teaching program must be judged), it is generally agreed that he is intellectually curious and exploratory, taking an active part in manipulating and reconstructing his environment. To paraphrase a number of current definitions:

> The creative child prefers to explore the unknown, rather than to conserve the already known; he prefers explaining facts in new ways, rather than continuing to rely on traditional well-established explanations; he indulges in adventuresome thinking and raises questions, rather than being content with things as they are.

With these preliminary comments setting the stage, we turn next to more pointed observations about the nature of creative thinking and how best to foster it – observations which to a large degree have determined the specific form and character of the research of the Berkeley group. As a first step, we will identify and discuss several widely-held misconceptions about creative functioning and its facilitation.

MISCONCEPTIONS CONCERNING THE NATURE OF CREATIVE THINKING

Perhaps the most commonly held belief concerning creative functioning is that the child will become creative of his own accord – by drawing on some heretofore latent potential – if only he can be placed in a stimulating yet permissive and nurturing environment. The erroneousness of this view is due not so much to a misreading of the nature of man as it is to an oversimplification of the means by which his latent capacities are to be realized. More specifically, an environment which is at once permissive and stimulating undoubtedly plays an important role in establishing the conditions necessary for creative output, but this is only part of the story. In order to take full advantage of an unrestrictive atmosphere, the child must first come to understand what constitutes creative ideas in the given situation and how he can achieve such ideas for himself. In short, he must learn how to think creatively. In this connection, a growing body of observational and empirical evidence (Hammer, 1961; Holt, 1964; Peel, 1960; Torrance, 1965) indicates that most school children are inadequately prepared for creative thinking. Corroborative data gathered by the Berkeley group at the fifth-grade and sixth-grade levels reveal a conspicuous inability among students to think of any ideas, much less clever or novel ones. Moreover, there is little understanding of the nature of originality and virtually no sense of planfulness which is necessary for prolonged effective work on a creative task. Not surprisingly then, one finds that the responses of most children to challenging tasks, even in a permissive environment, are by and large pedestrian and generally lacking in creative merit.

In order to think creatively the child must have at his disposal a repertoire of creative thinking skills. To name a few: the ability to recognize gaps in existing information; a facility at formulating relevant questions, and a sensitivity to the demands of the task so one can adequately judge the suitability of proposed ideas. The epitomy of creative thought is a sense of "disciplined abandonment." An array of cognitive skills such as those just mentioned constitutes the "discipline" component which must be present in concert with an accepting nurturing environment before the child can take full advantage of such freedom and "abandon" himself to the task at hand. Thus, seen in this perspective, the job of fostering creative thinking is a more complicated undertaking than suspected by many. It is more than merely providing a permissive atmosphere for the release of creative potential; it is more like teaching the child how to make use of his freedom to create.

If one must teach children to think creatively, then what kinds of pedagogical strategies would seem most appropriate? There are, no doubt, a number of reasonable answers to this question. For example, one approach – the traditional one – is to teach in such a manner that the child will come to act creatively in a number of different subject-matter disciplines. Another contrasting, yet complementary possibility is to teach for a number of cognitive skills fundamental to all creative thinking and then show the student how such generalized skills can be applied in specific subject matter areas. This latter strategy carries with it the implication of developing curriculum materials whose subject-matter is the creative thought process itself. As we shall see presently, it is this approach which is being explored by the Berkeley group.

Although there is considerable latitude with respect to the variety of reasonable approaches for promoting creative thinking in the classroom, there are limits. Some limits can be established based on our growing knowledge of the dynamics of the creative process. For example, it has been well established that creative thinking is largely dependent on cognitive operations and dispositions of a nonlogical character. Yet despite this, the notion as well as the practice persists of teaching for logic and for critical thinking as means for promoting creative functioning. As we shall see shortly, creative thinking, whether it be in science or in the humanities, involves a good deal more than rational and critical analysis, and to teach only for these functions to the exclusion of more imaginal, speculative processes is to fall short of fostering creative thought.

It has been further argued that teaching for scientific methods of thinking will establish a respect for the lawfulness and the simplicity of nature itself. Although the child may indeed infer from such exercises that nature is orderly and lawful, such an understanding may curiously enough act to fetter creative thought. More specifically, as currently taught, elementary school science, the humanities, and the social sciences as well, do not always reflect the way scientific investigation and creative scholarship actually happen. Creative contributions to any field of knowledge are made by wrestling with ambiguous conflicting facts and data with the overall aim of bringing conceptual order out of chaos. However, as represented in many current textbooks, the drama of the pursuit of such order and understanding is replaced by an oversimplified, tailored, and predigested view of a discipline, emphasizing what is already known and stressing the all-encompassing explanatory power of presently accepted theories or laws. Such a simple, clear-cut presentation which smooths over complexity and controversy is designed primarily to make for more efficient learning and retention, rather than to focus on the problem of how the child himself can learn to manipulate a given set of facts, to draw his own conclusions, or indeed to discover new facts, which is after all, one of the primary functions of the creative innovator. There is little question that such textbook presentations will make for the most rapid, untroubled assimilation and mastery of the material, but it should not be assumed that the approach which makes for the most efficient learning will also be most effective in fostering creative and productive thinking.

Moreover, in the way most subject matter has been traditionally taught, the student is rarely allowed to practice on problems which require innovative modes of thought for their solution. Consequently, he has no direct opportunity to learn what constitutes creative ways to manipulate data, nor to develop an intuitive sense for designing research, nor to ask questions which favor serendipity and maximize the chance for new insights. In short, the child does not acquire the cognitive skills which would be most helpful in preparing him for future productivity.

A specific case in point illustrating the lack of emphasis on thinking creatively about facts and information is the set of so-called "discussion questions" found at the end of many textbook chapters. As commonly used, these questions are not particularly helpful in promoting creative patterns of thought. Often the student is merely asked to think about these questions with little or no subsequent follow-up discussion of the adequacy of his answers. With minimal dialogue between the student and the teaching authority, the child rarely gets any systematic feedback as to what counts as merely an adequate answer as opposed to an ingenious, novel one. Even in those cases where the teacher finds time to comment on the quality of the pupil's essays, the feedback is inevitably delayed, thus reducing the effectiveness of the exercise. Because the child typically operates from question to question in what is essentially a "cognitive vacuum" with little opportunity to profit from his past efforts, he is unable to develop any internal standards of excellence against which he can evaluate subsequent ideas. In effect this means he is unable to modify his efforts in the direction of more creative and imaginative output.

Another feature of such discussion questions – their placement at the end of the chapter – strengthens the impression that these exercises are simply an appendage, added to the text as an afterthought rather than forming a central part of the learning experience itself. Moreover, their placement occurs at the psychologically inappropriate moment for stimulating productive ideas. By the end of the chapter much of the intellectual tension and drama build up as part of the content presentation will have dissipated, making it difficult to engage further the child's interest. In contrast, the placement of discussion questions should be coordinated with the presentation of the content itself, capitalizing on the student's immediate curiosity and momentarily aroused interest.

RECOMMENDATIONS FOR A CURRICULUM OF CREATIVE THINKING

Recognizing some of these pedagogical deficiencies, what can be recommended with regard to a curriculum for creative thought? The first recommendation is that such a curriculum feature repeated opportunities for the child to practice wrestling with complex challenging tasks of the kind that might, in a more complex form, engage the efforts of creative adults. Of course, such problems (which we shall call "creative-tasks-in-miniature") must be scaled down and simplified in many respects – in particular, reducing their dependence on technical or specific knowledge. Nevertheless, even with various modifications of this type such tasks could still retain the basic elements of any intellectually creative undertaking. A second recommendation is that certain broad guides and strategies for creative thinking be introduced as the child works on these tasks. Such guides would act as points of reference around which the child could organize his work and would impart an atmosphere of planfulness to the process of problem solution.

Next, what can be said more specifically about the nature of these "creative-tasks-in miniature" – their particular content and the composition of the accompanying rules and strategies for productive thought? It is sobering to realize that the creative innovators of the year 2000 (only 32 years hence) are to be found in the elementary schools of today. As these individuals stand on the threshold of the twenty-first century, they will be confronted by scientific revelations and social issues literally undreamt of today as well as a large share of old problems left unsolved by previous generations. To meet these challenges the innovator of tomorrow must create new social systems, new scientific techniques, tools, and theoretical structures. To accomplish this, he cannot rely on presently fashionable views as to what constitutes fruitful scholarship or good scientific method, nor can we afford to instill in him a blind devotion to currently fertile theories, since much of this theoretical apparatus is already approaching obsolescence. Here then is a compelling argument for the necessity of teaching broad strategies for creative thought and of encouraging cognitive flexibility and intellectual curiosity, rather than merely teaching for what we presently take to be true or useful.

What kinds of tasks could children practice on now to prepare them for their roles as the innovators of the future? Two kinds are suggested: the hypothetical problem and the enduring problem. The former type minimizes reliance on present facts and theories in obtaining a solution, while the latter type focuses on the deathless issues of human affairs which are certain to be the subject of creative reinterpretation and re-evaluation in any future age.

Typically in the case of the hypothetical problem certain artificial conditions are postulated which do not correspond to present realities, and the individual is called on to operate within these internally consistent but artificial systems. One problem of this type, which itself has become popular as a measure of creativity, is the consequence test: given certain conditions, such as a world in which time runs backwards, what are all the consequences? Such problems demand cognitive flexibility, imagination, and the ability to adopt as temporarily real a set of foreign and unfamiliar assumptions.

At Berkeley we are now drawing up plans for an experimental teaching apparatus which simulates a visit to a strange planet. The student is put in charge of the first exploratory expedition. His job, a highly creative and demanding one, is to come to understand and to explain the welter of strange events which he will encounter in this alien, unfamiliar world where the usual physical, biological, and social laws are suspended or greatly modified. Here the child is literally confronted with a hypothetical world which, although it is ordered and internally consistent, operates in ways that run counter to his previous experience. He must suspend his present expectancies about the world as he knows it long enough to create new explanations for strange and perplexing phenomena. Thus, by permitting the child to deal with hypothetical tasks whic call for the creation of new systems of ideas, independent of present facts and assumptions, he can hedge against the time when these facts will outlive their usefulness and will no longer serve as a basis for productive thought.

Granting the largely unknown course of future events, there are nevertheless certain enduring problems of man and society which are constantly reasserting themselves and no doubt will engage the attention of future creative thinkers. The young student of today can best be prepared to cope with such issues in adulthood by being presented now with repeated opportunities to deal with enduring themes and quandries: to search for new implications and to re-phrase them in contemporary terms. Many of the most suitable teaching examples are found in literature, such as the Greek myth of Sisyphus, dealing with man's reluctance to accept his fate; the tragedy of Frankenstein, reflecting the consequences of man's age-old dream of creating a more perfect being is his own image; or the mythical Icarus, personifying man's unflagging zeal to explore nature regardless of the consequences. Of course, it is understood that such issues cast in the form of teaching examples would have to be greatly simplified and made appropriate to a child's level of understanding. Notice, however, that such modifications are made easier by the fact that the traditional vehicles for the dramatization of these issues are themselves reasonably constant – the novel, the play, and the poem – providing already familiar media through which children can express their ideas. In this connection, we shall presently describe a "creative-task-in-miniature" in which the student creates a play dealing with yet another timeless and recurring theme – the plight of the wrongfully accused.

Finally, what are the sources on which one can draw in formulating a set of broad cognitive skills and strategies which could be taught now to prepare children for future productivity? Information concerning such generalized cognitive strategies can be gleaned not only from laboratory research dealing with complex problem solving and thinking but from anecdotal and life-history accounts of highly creative individuals as well. In the first instance there is a steadily accumulating body of knowledge from the psychological laboratory concerning the kinds of factors, both facilitative and inhibitory, which influence the course and quality of productive thought. One widely documented example is the deleterious effect of rigid mental set on effective problem solving (Luchins, 1942). Here, because the individual initially formulates or perceives a task in a narrow and inherently biasing way, he is effectively sealed off from attaining a solution. A number of appropriate thinking strategies incorporating these and other research findings can be formulated. For example, with reference to the present example of mental set, the student can be taught to view a problem broadly, to re-phrase questions in new ways, and to resist jumping to premature conclusions.

The other source of cognitive strategies comes from anecdotal descriptions of creative individuals at work (Ghiselin, 1952; Koestler, 1964). A number of insights into the fundamental nature of creative thought have grown out of such observations as well as the discovery of important phenomena associated with creativity, such as incubation and intuition. As an example, the famous and widely cited self-reports of Poincare, the French mathematician, have served to focus attention on the conceptually fertile notion of creativity as basically a process of recombining ideas in new and novel patterns. And of most relevance to our immediate concern, such observations have led to the elucidation of a number of specific problem-solving strategies, among them the technique of identifying the essential elements of a problem situation and deliberately juxtaposing them in various ways (Campbell, 1960).

By combining the experimental laboratory approach and its virtues of empirical validation with the rich and broadly inclusive self-reports of the highly creative thinker, one can formulate a set of rules and strategies which are pervasive enough to apply to a variety of creative tasks, yet simple enough to be readily understood by elementary-school children.

In summary of the first section of this paper it has been argued that before the student can take full advantage of a permissive stimulating atmosphere he must be taught how to think creatively. In developing a teaching program for creative thought the single most important pedagogical issue is the long-term social and personal usefulness of what is taught. It has been suggested that one reasonable teaching strategy, in light of this consideration, is to strengthen those cognitive skills basic to all creative thought in the context of complex yet meaningful problems which reflect the principal steps in the creative act. These observations, taken as a whole, have formed the broad guidelines for the research carried out by the Berkeley group over the past several years.

THE PRESENT RESEARCH

As a first step it was decided to develop a set of prototype instructional materials dealing with only one aspect of productive thought– creative problem-solving. If the particular teaching approach showed promise, then other programs would be designed focusing on yet other domains of creative thought such as creative innovation and creative expression.

After several years of intensive effort involving numerous pilot tryouts, successive revisions, and several full-scale experimental studies, the General Problem Solving Program was developed (Covington, Crutchfield, and Davies, 1966). The General Problem Solving Program (GPSP) consists of a series of sixteen "creative-tasks-in-miniature" designed for the fifth-grade and sixth-grade levels. These problem episodes act as a vehicle by which the student practices a number of broad rules and strategies concerned with various facets of effective problem solving, such as stating the problem in an open and unbiased fashion, reformulating the problem in terms of familiar metaphors and analogies, or developing a systematic plan for generating ideas.

Each lesson presents a mysterious occurrence or unexplicable happening which the student tries to explain, such as the puzzling behavior of a group of aquanauts during a deep-sea dive. This general theme – the mysterious occurrence – was chosen because it encompasses in a natural and uncontrived manner many of the fundamental aspects of the problem-solving process. In addition, such a theme is neutral with respect to orthodox curriculum materials. This makes for a broader transfer of what is taught than might otherwise be the case if the principles and concepts had been associated primarily with a specific content area. Finally, children exhibit a keen interest in working on problem situations featuring a mystery-detection theme.

Each lesson is presented in booklet form and is cast in a self-instructional linear format. (For a theoretical discussion of the potential uses of programmed instruction for fostering higher-order thought processes, see Crutchfield and Covington, 1965.) Briefly, the child works through the lesson in a step-by-step sequence by himself and at his own pace. As the problem unfolds page-by-page, with the advent of new facts and clues, the student is called on to reformulate the problem in his own words, to list questions he considers crucial to the solution or to devise a plan of action which he intends to follow. The student receives immediate guidance in his efforts on successive pages of the booklet. This guidance typically takes the form of presenting a range of ideas, questions or courses of action which the student might have thought of in the given situation. Here the emphasis is on encouraging appropriate diversity in the student's thoughts by showing him within wide limits what constitutes valuable ideas, crucial questions and fruitful ways to formulate problems.

The GPSP is also designed to promote beneficial attitudes toward productive thinking. One of the basic strategies is to increase the child's experience in coping successfully with thought problems. Each problem episode is sequenced so that as the student works through the booklet he is exposed progressively to more clues and hints. Thus, each child, independent of his initial capacities or personal reservations about thinking, will at some point come to discover the solution for himself. In addition there is a systematic attempt to foster, by means of appropriate teaching examples, a number of attitudes which favor effective problem solving, such as open-mindedness, persistence, and suspension of premature judgment.

A story-line is maintained throughout the GPSP to supply a sense of continuity, and perhaps most importantly, to provide the reader with a set of identification models. The story concerns two school children, Jim and Lila (brother and sister), whose extra-curricular pastime is the exploration of mysterious situations which arise in and around their hometown. Jim and Lila are assisted – and when the occasion demands, assiduously directed step-by-step – through these adventures by their uncle. The uncle in addition to being a high-school science teacher also "moonlights" as a detective.

The identification-model technique is intended to introduce the reader gradually to the difficult and often frustrating process of becoming a more effective thinker. For example, the roles of Jim and Lila reflect the vicissitudes associated with all complex problem solving – the long discouraging periods of intense effort with little apparent progress, the inevitable setbacks, and finally, the intellectual elation of discovering a solution. Through Jim and Lila the reader can experience all this vicariously without being plunged immediately into the situation. After the first few lessons, however, the reader is gradually drawn into the various problem-solving activities by being requested to think of and to record his own ideas in concert with the efforts of Jim and Lila – first the student generating his own ideas or questions, then Jim and Lila responding with theirs. The feedback examples provided for the reader are presented as Jim's and Lila's ideas.

Against this backdrop of transitory day-to-day successes and failures, a long-range change takes place in Jim and Lila. They are depicted as overcoming initial handicaps of reticence, apathy, and negativism toward thinking until ultimately, but not without an occasional setback, they become reasonably comfortable with their own thought processes, much more enthusiastic about tackling problem situations, and more confident in their own abilities to cope with them. It is hoped that this subtle but perceptible change in Jim and Lila will foster in the reader a sense of his own progressive improvement. At the same time it is intended that as many readers as possible will finally come to surpass Jim and Lila in problem-solving proficiency and will outgrow any earlier dependence on them as a source of ideas and inspiration.

The inclusion of the uncle provides Jim and Lila, and presumably the reader as well, with a benevolent authority figure and confidant who not only nurtures and encourages the nascent attempts of these tyro-thinkers but who also shows them how to think for themselves by means of appropriate rules and strategies. Additionally, the uncle stands as a valued model personifying an enthusiasm for intellectual exploration.

Each lesson is presented in an illustrated format. Such a dominantly visual presentation not only increases the student's interest in the materials, but makes it easier for him to follow the necessarily complicated dialogue and thought sequences of the story characters. The lessons contain an average of 40 pages and are designed to be presented at the rate of one lesson per day. The average time taken to complete a lesson is approximately 35 minutes.

Several separate studies (summarized in Covington and Crutchfield, 1965) have been carried out using various preliminary editions of the GPSP which to date have involved a total of some 481 fifth-grade and sixth-grade school children from the San Francisco Bay Area, of whom 267 were given the training materials and the remaining 214 were used as controls. The basic design of these studies is essentially the same. All students are first administered an extensive pretest battery (six hours in length in the latest study) which consists of a number of tests of creative thinking (Torrance, 1965), tests of problem-solving ability (Covington, 1966c), inventories designed to measure the child's attitudes toward problem solving and thinking (Olton, et. al., 1967), and various tests of school achievement and scholastic ability. These data are used as a basis for the initial matching of the classrooms. Individual classrooms are matched in pairs; one classroom from the pair is then assigned at random to the instructed condition and the other to a control condition. In each of the various studies to date, the instructed groups have always been administered the GPSP. In contrast, a number of different control conditions have been used to test varicus hypotheses about the nature of the training effect. The typical case, however, and the one on which the present analysis is based, calls for the administration of a similar but shorter set of self-instructional materials whose content is unrelated to problem solving. The purpose of the control program is to insure a sense of involvement and participation on the part of the control children and to make certain that they are well acquainted with the self-instructional format. Following the training period all classes are administered a posttest battery which includes a repeat of the attitude inventories and the tests of problem-solving ability along with parallel forms of the tests of creative thinking.

Several examples will suffice to illustrate the general types of tests included in the pretest and posttest batteries – their content, style and degree of similarity to the training material. One type reflects the more traditional tests of problem solving. In these instances, because of a large number of restrictions placed on the task, there is only a limited set of answers or principal solutions which will satisfy the problem in its entirety. Here an emphasis is placed both on convergent and divergent thinking. The student is called on to diverge in his search for appropriate ideas – opening up a host of possibilities – but at the same time to converge, focusing on the few most suitable ideas. A representative sample is the X-ray problem (a modification of the classic problem used by Karl Duncker). The child is required to invent a method to kill a tumor deep inside a body by using an X-ray but without harming the surrounding healthy tissue. The primary constraint is that if the X-ray is too strong it will kill both the healthy tissue and the tumor. If, on the other hand, the X-ray is made too weak, it will not harm the good tissue, but neither will it kill the tumor. Incidentally, this problem in addition to possessing a high degree of curriculum relevance is quite unlike anything encountered in the training lessons in terms of content or theme.

The other main type of problem included in the criterion batteries is the Minnesota Tests of Creative Thinking (Torrance, 1965). Such problems are quite open-ended with few constraints. This allows for a number of solution-ideas, no one of which can be judged as the most suitable. One example featuring a product-improvement theme requires the child to think of all the ingenious clever ways to make a toy dog more fun to play with.

As to the results: in general, the outcomes of the various studies employing the GPSP have been highly consistent. The performance of the instructed children is markedly superior to that of the control children both on the tests of problem-solving ability and on the tests of creative thinking. (For a detailed report of findings as well as a discussion of scoring procedures, see Olton, et. el., 1967.)

In order to gain a clearer picture of the over-all magnitude of this training effect, consider several performance indices such as total ideas generated, total quality of ideas, mean quality per idea and the incidence of principal solutions. Composite scores for each of these indices are computed for every student by summing over his performance on a number of problem-solving tests in the posttest battery. In the same manner a similar set of composite scores is obtained for the tests of creative thinking. On comparing the mean values for the instructed and control groups on any of these composite indices it is found that the instructed groups are invariably superior. In most cases the magnitude of these differences is absolutely large and beyond mere statistical significance. This is illustrated by the fact that when comparing frequency distributions for each composite index, the fiftieth percentile of the instructed groups typically falls around the seventy-fifth percentile for the control groups.

Another way to gauge the degree of educational significance of a training effect is to compare the level of performance of instructed and control children of differing IQ levels. In this connection it has been found (Covington, 1965; Covington, 1966b) that the mean performance of instructed children with IQ's below 99 (mean IQ = 91) is on a par with the mean performance of control children whose IQ's fall between 100 and 115 (mean IQ = 107). This indicates that the administration of the GPSP makes for a substantial boost in the performance of low IQ children over a wide variety of test problems, many of which are curriculum relevant.

To determine how long these training effects persist, a follow-up test battery was given five months after the administration of the posttest in two different studies. These follow-ups included approximately 80 percent of the fifth-grade children from the original samples who by this time were in sixth-grade classrooms with new teachers. Wherever possible every precaution was taken to insure that the students did not recognize the tests as a part of the earlier work. For example, the children recorded their ideas on regular school paper, rather than in a standard booklet of the kind used in the original study. All the tests were administered by the teacher herself, and whenever possible were introduced as part of the regular classroom work.

In the case of tests of problem-solving ability, the instructed children continued to surpass their control counterparts on the various performance indices, and although the margin of differences had diminished somewhat as compared to the magnitude of effects found at the time of the original posttest, the differences were nevertheless found to be statistically significant. The picture is not as consistent for the tests of creative thinking. In one sample with a total of 108 children, the instructed group was superior to the control group, whereas in another slightly larger sample these differences washed out. It appears that after a five-month interval the training effects are at best marginal in the case of creative thinking tests. This is not surprising when it is realized that as a group these tests represent a type of problem not directly trained for in the GPSP. Consequently they would be more likely to reflect a greater diminution in training effect than would the tests of problem-solving ability which are more directly amenable to the kinds of strategies taught for originally.

ISSUES OF VALUE AND JUSTIFICATION

One of the most fundamental questions raised by the present research is whether programs designed to promote creative thinking are actually needed, seen in the perspective of a child's total educational career. No matter how beneficial a given teaching program may be, it is likely that untutored children (the so-called controls) will sooner or later catch up, simply through the normal process of intellectual maturation and accumulating experience. In the present data the sizable reduction in the magnitude of the training effect five months after the administration of the GPSP supports this contention. Of course, it can be argued that more permanent – and in this sense more meaningful – changes will occur only on an accumulating basis, and that what is needed is a program covering a whole school year or even longer, consisting of interlocking, coordinated curriculum units which build on one another. It remains a moot point, without empirical support, yet it would seem likely that a long-range program of increasing scope and complexity designed to stimulate intellectual growth would allow the instructed child not only to maintain a performance superiority, but actually to increase that margin as the program proceeds. While it is true that the trained and the untrained student alike enjoy a natural expansion of intellectual capacities, there is grave doubt that the untutored children would ever make use of their burgeoning capabilities to the same extent as would children who received systematic long-term training. And, as we have seen, since most elementary-school children make scant use of their present capacities for creative thought, there is little reason to believe that these same children at a later age would spontaneously draw on their capacities to any greater degree.

A related point concerns the developmental changes in attitudes and values which favor the exercise of creative thought. While it is quite clear that sheer proficiency in various cognitive skills such as question asking or idea generation is a function of age, there is as yet no evidence that the relevant attitudes and values increase as the child grows older. Thus, even though untutored children may in time overtake the trained children in terms of sheer proficiency, they may be markedly deficient in the very attitudinal dispositions necessary to put such skills to meaningful use. In this connection, creativity training can be designed to provide the student with experiences he is unlikely to receive anywhere else – the challenge of working on a complex but meaningful problem or the sense of satisfaction at discovering a solution – experiences which may affect the child's emerging set of values and attitudes regarding creating functioning. Such experiences, if they occur at decisive points in the individual's development, especially in childhood, may be sufficient to bring him permanently past a critical threshold for actualization of his creative potential. As a matter of fact, we may find in the last analysis that the key to the unlocking of creative potential is to strengthen sets of attitudes and values which predispose the student to undertake creative tasks in the first place, rather than to start by increasing his sheer competence for performance by means of "skill training."

FUTURE RESEARCH PLANS

Research plans for the immediate future take two main directions. The first of these involves further development and expansion of the GPSP. In one instance a set of supplementary exercises is being developed to provide additional practice on the various skills and strategies taught in the lessons. These exercises are primarily for remedial purposes and can be used when needed, depending on the rate of progress of the individual learner. Another project concerns the development of a set of "curriculum-link-units," that is, lessons designed to illustrate how the student can apply the skills and strategies taught for in the GPSP to his actual school work in science and the social studies.

The other main direction of research involves the development of new self-instructional teaching units which introduce the student to domains of creative thinking other than problem solving. One obviously important area is that of creative understanding. Here, typically, one is confronted with a series of complex events or occurrences which in the initial phases of investigation often appear to be unrelated. The task is one of discovering meaningful relationships between such events and in some instances of predicting how these events would covary under various conditions. In the case of problem solving defined in a more traditional sense, the search is typically of a more limited nature – that of discovering a single workable solution-idea which satisfies a certain requirement, but often without the necessity of understanding why it works. Such a pragmatic emphasis tends to make for closure once a solution is obtained. On the other hand, creative understanding not only encompasses the discovery of workable ideas, but more often involves an ongoing procedure of checking and re-checking the validity of the proposed solutions against an accumulating set of facts and information.

Another area of creative output, perhaps the one most generally assumed when reference is made to creativity, is that of creative innovation. Here the individual invents or creates new problems, new systems of thought, or new products which did not exist previously. One example of a prototype training unit presently being developed involves the creation of a school play. The story-line concerns some malicious damage done to a barn owned by a local farmer and the farmer's unwarrented accusation that a certain boy is responsible. The boy, who is innocent, is able to convince his classmates of this fact, but is unable to prove it to a degree that satisfies the farmer and the rest of the adult community. Understandably the children react with indignation, but at the same time they realize their essential helplessness at being unable to defend adequately their interests in an adult world. While some of the children make plans to track down the real culprit, others decide to express the group feeling of ineffectuality and frustration by means of a play. Parenthetically, such a plot can be used to illustrate that there are other alternatives to the expression of one's emotions besides direct aggression and that intense feelings can serve as a basis for the development of useful and satisfying products. The particular theme – the plight of the wrongfully accused – was chosen because it sounds a note of high drama and arouses immediate interest and sympathy in children who typically have a keenly developed sense of fair play. Moreover it represents – even when placed in such a simplified context – one of the enduring issues of man and society which demands new solutions from each succeeding generation.

The same general pedagogical devices are used here as in the GPSP. The student is led to develop the play in a step-by-step sequence, while thinking of his own ideas and being provided in turn with appropriate feedback and redirection of his efforts. In the early lessons the child works through sequences dealing with the creation of the basic plot. The student is encouraged with the assistance of identification models in the story to generate ideas which seize on the malicious incident as a point of departure from which to explore the more general implications of being wrongfully accused. This is in contrast to the more pedestrian and common-place approaches of simply dramatizing the boy's innocence or holding the farmer up to ridicule. Basic to the success of this teaching sequence is the introduction of metaphorical and analogical modes of thinking. Once such concepts are grasped by the student he is able to go far beyond a strictly literal interpretation of a situation and can generate a virtually limitless set of implications.

After the outlines of a clever insightful plot are secured, the student explores a number of detailed matters, all of which are important to the final product. In one lesson he creates ideas for the most effective scenery and staging, given a limited amount of material to work with; in another he is called on to create segments of the dialogue. The last lesson in the series simulates the public reaction to the play which represents a final confirmation of all the preceding activities. Various members of the audience including the farmer and the parents of the accused boy react to the play. There is general recognition that it was worthwhile in its own right as well as being helpful in clearing up a community misunderstanding.

FINAL PARADOXES AND DILEMMAS

Inherent in any attempt to teach for creative functioning are a number of dilemmas and paradoxes which give rise to both methodological and theoretical difficulties. In conclusion we will consider one recurring dilemma, not because it is necessarily representative or even the most fundamental, but rather because it indicates something of the extraordinary challenges encountered in developing teaching materials for fostering complex cognitive processes. Basic to the act of teaching is a guidance function. To a greater or lesser degree the student's behavior is necessarily guided, shaped, and redirected by a teaching authority. The student never has complete freedom. He is exposed only to certain learning experiences, in certain teacher-determined sequences. He is encouraged to talk about these experiences in only certain ways, using pre-arranged kinds of terminology. If we teach for creative functioning, then the child will inevitably be subject to some form of guidance. To the extent that an educational theory stresses the "discipline component" of creative thought, any forthcoming curricula will be more or less restrictive and predetermined. How then can we reconcile these procrustean-like features of teaching with the fact that in the last analysis creativity is primarily characterized by imaginal freedom and a spontaneous individualization of thought? Perhaps the answer, most simply put, is to teach for structured spontaneity. One technique adopted for this purpose and employed in our current work is to give the student direct practice in discriminating among ideas which are both unique and appropriate to a given task and ideas which are merely bizzare. In such a fashion the student learns to limit the reaches of his spontaneity. Another technique which requires a relatively long term training program involves comprehensive guidance in the early phases of learning. But, as the child becomes more proficient in the various skills and strategies, the rigid guidance of the program is gradually reduced. In this manner the student comes ultimately to rely on his own resources and initiative with only occasional redirection from the program. As the student becomes more self-sufficient he can experience yet another dimension of intellectual independence – that of determining for himself which of several creative tasks he will work on from among a number of alternatives. For example, in the case of the creative drama unit just described, the child could be given the option of either writing dialogue or of developing the scenery and staging, once he had completed the introductory units.

All the foregoing implies a complex catering to the individual differences of the learner; first, an intensive close-knit and personalized guidance of the child in the early phases of learning and later, encouragement of self-direction and self-determination. In both cases the training environment must be capable of selecting, sequencing and coordinating a great many combinations of material. This immediately suggests computer-assisted teaching. The self-instructional format of the current and proposed teaching programs is admirably suited to a computer-assisted operation. Indeed, another facet of the work of the Berkeley group is that of adapting the GPSP for a computer presentation, thus by-passing the cumbersome booklet format with its limited feedback capabilities. With computer assistance the child can be guided through a much more individualized sequence of learning, dictated by his initial level of competency, his particular cognitive style, and his individual rate of progress. Conceptually, it may appear that computers and programmed instruction are antithetical, perhaps even perversely so, to the task of promoting creative thinking. However, such an unlikely juxtaposition points to the kinds of imaginative responses which the educational community must make to the challenge of educating individuals in an increasingly complex and demanding world.

FOOTNOTE

1. The research reported in this paper was supported by a grant from the Carnegie Corporation of New York under the direction of the author and Dr. Richard S. Crutchfield. This paper

covers research conducted through 1965. Summaries of our more recent research activities can be found in Olton, et. al., (1967) and Olton and Crutchfield, in press.

REFERENCES

Barron, F. The relationship of ego diffusion to creative perception. In C. W. Taylor (Editor), Widening horizons in creativity. New York: Wiley, 1964. Pp 80-86.

Campbell, D. T. Blind variation and selective retention in creative thought as in other knowledge processes. Psychological Review, 1960, 67, 380-400.

Covington, M. V. The effectiveness of training for problem-solving proficiency and creative thinking as a function of differing ability levels among children. Paper presented at the meeting of the Western Psychological Association, Honolulu, June 1965.

Covington, M. V. A childhood attitude inventory for problem solving. Journal of Education Measurement, 1966, 3, 234. (a)

Covington, M. V. Programmed instruction and the intellectually gifted: Experiments in productive thinking. Paper presented at the Symposium on Programmed Instruction and the Gifted Child, International Convention of the Council for Exceptional Children, Toronto, April 1966. (b)

Covington, M. V., and Crutchfield, R. S. Experiments in the use of programmed instruction for the facilitation of creative problem solving. Programmed Instruction, 1965, 4.

Covington, M. V., Crutchfield, R. S., and Davies, Lillian B. The Productive Thinking Program. Series One: General problem solving. Berkeley: Brazelton Printing Company, 1966.

Crutchfield, R. S. Independent thought in a conformist world. In S. M. Farber and R. H. L. Wilson (Editors), Control of the mind. Part 2. Man and civilization: Conflict and creativity. New York: McGraw-Hill, 1963. Pp. 208-228.

Crutchfield, R. S., and Covington, H. V. Programmed instruction and creativity. Programmed Instruction, 1965, 4.

Duncker, D. On problem-solving. Translation by Lynne S. Lees. Psychological Monographs, 1945, 58, (5, Whole No. 270).

Ghiselin, B. The creative process: A symposium. New York: New American Library, 1952.

Hammer, E. F. Creativity: An exploratory investigation of the personalities of gifted adolescent artists. New York: Random House, 1961.

Holt, J. How children fail. New York: Pitman, 1964.

Koestler, A. The creative act. New York: Macmillan, 1964.

Luchins A. S. Mechanization in problem solving: The effect of Einstellung. Psychological Monographs, 1942, 54, (6, Whole No. 248).

MacKinnon, D. W. Personality correlates of creativity. In Mary Jane Aschner and C. E. Bish (Editors), Productive thinking in education. Washington, D. C.: National Educational Association, 1965. Pp. 159-171.

Mednick, S. A. The associative basis of the creative process. Psychological Review, 1962, 69, 220-232.

Mendelsohn, G. A., and Griswold, Barbara B. Differential use of incidental stimuli in problem solving as a function of creativity. Journal of Abnormal and Social Psychology, 1964, 68, 431-436.

Myers, R. E., and Torrance, E. P. Invitations to thinking and doing. New York: Ginn, 1964.

Olton, R. M., Wardrop, J., Covington, M. V., Goodwin, W., Crutchfield, R. S., Klausmeier, H. J., Ronda, T. The development of productive thinking skills in fifth-grade children. Technical Report, Research and Development Center for Cognitive Learning. Madison, Wisconsin: The University of Wisconsin, 1967.

Olton, R. M., and Crutchfield, R. S. Developing the skills of productive thinking. To appear in: P. Mussen, J. Langer and M. V. Covington (Editors), New Directions in Developmental Psychology. New York: Holt, Rinehart and Winston, in press.

Osborn, A. F. Applied imagination. New York: Scribner's, 1953.

Parnes, S. J. A self-instructional course in creative problem solving. Unpublished manuscript, State University of New York at Buffalo, 1965.

Peel, E. A. The pupil's thinking. London: Oldbourne, 1960.

Suchman, J. R. Inquiry training: Building skills for autonomous discovery. Merrill-Palmer Quarterly, 1961, 7, 147-169.

Torrance, E. P. Rewarding creative behavior. Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

Upton, A., and Sampson, A. V. Creative analysis. Whittier, California: Whittier College Press, 1961.

Measuring Motivation in Culturally Disadvantaged School Children

Richard deCharms
Washington University
St. Louis, Missouri

Virginia Carpenter
Lindenwood College
St. Charles, Missouri

A COLLEAGUE of mine recently spent a semester in seventh grade (Smith and Geoffrey, 1965). In a school serving a culturally disadvantaged neighborhood, he tried to attend class every day and played the role of a nonparticipant observer. His account of this experience is very discouraging to a person who is primarily interested in human motivation. Take, for instance, one simple incident:

> Harry had been over at the confectionery across the street, and had come out with the rest of the boys on Tuesday, when a fight began on the playground. There had been considerable excitement and Mr. Inman (the principal) was trying to calm everyone down. Harry had left his pen over in the confectionery and wanted to go back across the street to get it. Mr. Inman had told him to wait for a few minutes. A little while later, he asked the question again and was told to wait. Then, or on the third time he had asked and was turned down and told not to bother Mr. Inman, Harry walked off and made a muttered and indecent comment. This was the end of it for him.

This commonplace incident in itself is minor and doesn't carry the impact of the whole report where such incidents cumulate. It can be taken, however, as illustrative. The utter futility of the situation is striking. This and other incidents make it clear that the children feel powerless to control their own destinies. As a result of their futile attempts to assert their initiative in inappropriate ways, they create an environment where even the teachers and administrators are forced to act in ways that are not of their choosing. The result is a constant struggle between pupils and the school staff to obtain the upper hand. Teachers search for ways to enhance their power to control the behavior of pupils, and the children use every means at their command to subvert the authority of the school, the last-resort being hostility through passive dependency. Both parties to the situation are ultimately pawns in the hands of the powers beyond their control.

Two more or less disconnected things stand out for me in approaching the problem of motivation in school children. First, the tangled web of influences will not yield to any miraculous new insights or techniques. The Gordian knot is real and will only yield very slowly to modest and patient efforts to unravel one or two threads at a time. In this age of crash programs, imminent breakthroughs, etc., motivation is an important concept to consider, but we must not expect it to solve all problems.

The second thing that strikes me is that one of the loose strands that may have potential for unraveling both practical and theoretical aspects of the knot is the notion of powerlessness, which is, after all, a motivational concept.

What I have to report then, is no breakthrough, but rather a modest beginning that attempts to discover whether certain aspects of motivation can be measured in a population of culturally disadvantaged children, and whether some of the motivational concepts akin to powerlessness can be used.

Actually we will discuss three different but related things. First, I will present some notions about powerlessness and what I will call the "Origin-Pawn" variable and discuss very briefly some relevant laboratory studies that we have completed. Second, I will shift to a discussion of achievement motivation and some of the conceptual and scaling difficulties that it presents. This is necessary in order to introduce a third section dealing with data on the achievement motive in culturally disadvantaged children. These three sections may seem a bit disjointed. What holds them together is the relationship between the concepts to be presented in section one and the achievement motive. At the present stage of the research there is little to report on these concepts except laboratory studies, but we do have evidence on achievement motivation from culturally disadvantaged children. As a result, I have more or less arbitrarily put the two together with a discussion of some theoretical problems with the measurement of achievement motivation.

POWERLESSNESS AND THE ORIGIN-PAWN VARIABLE

In the anecdote related above, Harry was frustrated by his powerlessness to influence an important aspect of his life vis-à-vis a school official. He was a pawn to forces beyond his control. In much of his experience in life, and especially in school, he is, in fact, and is expected to be, a pawn. Learning seems to demand that he react to outside influences in the school, and one way to insure that he will react in the proper manner is to reduce his freedom to a minimum, to force him to study, to press him into the mold expected by the school. Techniques designed to do this demand that the teacher and the school have most of the power and that power be taken away from the student.

Motivationally there is something wrong with this picture. The student, caught up in a power field, will do everything he can do to break out of it, and when he leaves school, he will be out of it. Is the goal of teaching to produce certain responses in the school that are almost certain to be a function of the specific power situation and, therefore, not transferred to non-school situations? Or is there a broader goal?

We would like to distinguish between the driven or power-induced school behavior, and behavior motivated from within the student that will carry over to extra-school behavior. The belief that the latter is possible reduces to a naive assumption that we all make in our everyday interaction. The assumption is that each individual has some autonomy and exercises it by originating his own behavior. Let me use the one word "Origin" to connote a person who is perceived to be originating his own behavior. A person is not always an "Origin." Often we perceive his behavior as stemming from a force field so that he is coerced to behave in a certain way by an external source. In such cases, I shall refer to him as a "Pawn" or as acting like a "Pawn." It should be clear that pawn-type behavior is induced by lack of power on the part of the pawn. Origin type behavior implies a feeling of power, although it is not entirely complementary in this respect to pawn-type behavior. Pawn behavior implies powerlessness, but origin behavior implies freedom of choice, not necessarily the ability to wield power over others.

Having sketched the "Origin-Pawn" dimension, let us note that it is essentially a phenomenological given in the perception of behavior. As such, it may be a myth or a mental way station, to use Skinner's railroading metaphor. To extend the metaphor, we might say that the main line of behaviorism runs nowhere near this way station, and efficient operation of the line would suggest that the station should be dropped. In fact, scientific psychology has been at pains to discredit the assumption that a human being can be the ultimate origin of some of his behavior.

Behaviorism, suffering under the strong influence of operationism and physicalism, has done its best to reduce the concept of motivation in psychology to something of the order of "the study of the causes of behavior." The trouble with this approach is that trying to elucidate the concept of motivation by means of the concept of cause is like trying to understand the behavior of a human being using concepts found to be inadequate to explain the behavior of billiard balls. The precept that all scientific statements must be based ultimately on statements about physical operations has great merit, but the hidden implication that explanations of physical motion (including behavior) are reducible to the laws of mechanics has seriously retarded the development of theories of human motivation.

Stripped of its hidden assumptions, the neo-behavioristic approach to motivation as the cause of behavior is unacceptable to anyone and has never been seriously held by anyone. It assumes that the human being is like a billiard ball and his behavior, like that of the ball, is entirely the result of external sources forcing him to behave. Another object hits him and steers him in a certain direction – an explanation that would account for certain trivial behaviors on the football field perhaps – or an external noxious stimulus repels him like one negative charge repels another, or an internal condition (e.g., drive-stimulus) produced by external circumstances (e.g., food deprivation) drives him in a certain direction.

Confronted with the phenomena of motivated behavior which in this framework appears akin to a sudden inexplicable movement of a billiard ball resting alone on a perfectly level table, psychology has ingeniously devised many schemes to explain this untoward behavior in terms of some mechanisms involving a causal chain of events and ultimately reducing to some undetected source of physical energy closely connected with, or actually residing in, an impelling stimulus (e.g., the cue ball). The cue ball, or impelling stimulus, has been variously described as a need or drive (Hull, 1943), a drive-stimulus (later Hull and Miller), incremental stimulation (Mowrer, 1960), or an affective stimulus (Young, 1961; McClelland, 1951).

In the face of all this theorizing based on what Gilbert Ryle (1949) would call a para-mechanical myth, it seems to me that Skinner has done the most consistent thing for a strictly empirically based "science of behavior," – namely to shun the notions of causality and motivation altogether. Although I suspect that he may have led some to hide their motivational and causal notions under the concept of reinforcement, his strict adherence to the definition of a reinforcer as something that increases the probability of a response does, in fact, rid the science of behavior of all of these irritating problems.

By now, I hope to have laid the groundwork so that two conceptions of man as a motivated being are clear. The first is the one we use in our everyday dealings with human beings: namely that the cause or reason for a person's behavior is sought first within him. More often than not, he is the origin of his behavior and is to be held responsible for it. The second is the explanation from scientific psychology that seeks a physical source external to the person, or at least a stimulus within him, that drives him. This latter notion has never been satisfactory in accounting for the phenomena of human motivation and has, through its emphasis on demonstrating a physical basis for all behavior, led to a psychology of learning and motivation that conceives of the animal and the human being as a pawn to be manipulated by reinforcement schedules.

Accepting at face value the psychological fact that people perceive themselves and others as origins in most instances is, for me, the starting point of a truly psychological approach to motivation, and demands a level of analysis quite different from the causal chain analysis that attempts to reduce the explanation to physical events. I have tried, therefore, to build on the assumption that the origin-pawn dimension makes a difference that is measurable in the behavior of a subject.

The origin-pawn dimension has its roots in the concept of locus of causality used by Heider (1958). His emphasis is on the perception of causality, and the result is that the concept is applied in the area of person perception. In perceiving another person's behavior, we do attempt to assess whether the locus of causality is internal or external to the person, and this distinction affects our perception of the person, the characteristics that we attribute to him, and the way we behave toward him.

Several person perception studies have demonstrated the significance of attribution of locus of causality (Thibaut and Riecken, 1955; Pepitone, 1958). In our own research, we have demonstrated (deCharms, Carpenter, and Kuperman, 1965) that subjects clearly do use the origin-pawn variable in their perceptions of others. We found that the perception of a person as an origin or as a pawn is a function of several variables, such as whether some agent is attempting to influence the person's behavior or not. Origin or Pawn perceptions are a function of characteristics of the influence agent, such as whether it is a large group, a small group, or an individual, and whether it is liked or disliked by the person. Most interesting, however, was the finding that a personality characteristic of the subject strongly affected his perception of another person along the origin-pawn dimension. Subjects were measured on a scale developed by Rotter, Seeman, and Liverant (1962) that attempts to measure the general feeling of powerlessness or efficacy of the person. Since Rotter uses reinforcement terminology, he calls the scale a measure of control of reinforcements and names it the I-E scale standing for internal or external control. There is a subtle difference between this concept of control of reinforcements and our O-P dimension, but for now the similarities may be noted. Our data showed that subjects who reported on the Rotter I-E scale that they felt in control of their sources of reinforcement perceived other people more as Origins, whereas subjects at the other end of the I-E scale perceived other people more as Pawns.

Since our primary interest is motivation, we next asked the question: "What effect does feeling like an Origin, as opposed to feeling like a Pawn, have on the person's own behavior?" With this question we moved from the perception of others to the perception of self as an origin or a pawn. Two laboratory studies (Kuperman, 1966; deCharms, Dougherty, and Wurtz, 1965) were devised that attempted experimentally to induce the person to feel like a pawn in one situation and like an origin in another, and to measure the effects on their behavior. The results confirmed the influence of the origin-pawn variable as manipulated in the experiment. Subjects building a tinker toy model under origin conditions that attempted to induce a free atmosphere, as compared to building a similar model under pawn conditions a) liked the origin model best, b) became more involved, c) chose to continue the origin model when interrupted, d) completed the origin model more elegantly, and e) recalled the nonsense name given the origin model more frequently a month after the experiment. Apparently, the origin-pawn dimension can be manipulated in a way that has strong effects on behavior. The Rotter I-E scale was administered to these subjects also, but was not related to any of the dependent variables, either directly or in interaction with the origin-pawn manipulation. Evidently, this questionnaire measure relates to other questionnaire responses as in the previous study of person perception, but doesn't predict behavior related to a task situation, at least for the male subjects used here.

ACHIEVEMENT MOTIVATION

So far our work on the Origin-Pawn variable has not ventured very far out of the laboratory, and we have very little empirical evidence about the variable in culturally disadvantaged school children. What we do have, however, is preliminary evidence about the measure of achievement orientations in this population, and some evidence that one of the aspects of achievement orientation involves acting like an origin. Let us be clear that I am not equating acting like an origin with having high n Achievement. What I am doing is looking for a relationship between the two concepts, and since I am just launching a project to attempt to measure the O-P variable and n Achievement in schools, and already have data on n Achievement, I will report the data in hand on n Achievement and speculate about the O-P variable. The connection is closer than may be apparent.

McClelland (1961) has spelled out several aspects of what he called the achievement syndrome, i. e., characteristics of people with high achievement motivation. One of the most important aspects is that the achievement motivated person likes to take personal responsibility for his actions. He likes to be in a position to act like an origin and avoids situations where he is a pawn.

Let us shift our emphasis, then, to the measurement of n Achievement in culturally disadvantaged children. First, let us be very clear about one thing. The word "achievement" is a loaded term, especially in educational circles. A culture such as ours, that has been nurtured on Horatio Alger stories and McGuffey readers, tends to perceive achievement motivation as an unmitigated virtue if not a blessing. If we pull back from this conception and ask what aspect of achievement motivation we can actually measure, we may be disappointed. The measure to be used was developed about fifteen years ago by McClelland and his associates at Wesleyan University, and the subsequent years of research have told us much about the characteristics of people who score high on the measure. The most appropriate single word to describe the person with strong achievement motivation as measured by this technique is that he is an "entrepreneur," typically following a business career. Top creative scientists are not particularly high on n Achievement, nor are academicians. If you want to succeed in business, however, you'd better have high n Achievement.

The achievement syndrome has three major aspects: 1) personal responsibility, 2) risk-taking strategies, and 3) the use of feedback. The entrepreneurial spirit of achievement emphasizes self-reliance, the taking of calculated risks, careful planning and checking of progress with constant emphasis on the skillful use of one's abilities. These, then, are the dependent variables that are related to the measure of achievement motivation.

What is the measure itself? Actually, even after years of research the original form of the measure has not been improved very much and is essentially a very crude measure. Originally intended as a measure of individual differences, it yields at best a very rough ordinal scale that is most useful when collapsed into a high-low dichotomy, or at most a high-medium-low trichotomy.

Before discussing the weaknesses of the measure, let us sketch the procedure for obtaining n Achievement scores. The technique involves collecting thought samples from subjects under standardized conditions by showing a group of subjects several pictures, one by one, for a short period of time and asking them to write a creative story about each in turn. The stories are content analyzed by trained coders for instances of Achievement Imagery. Stories in which one of the characters is concerned over obtaining success in competition with a recognized standard of excellence are said to contain achievement imagery. The construct validity of the measure is attested by years of research although test-re-test reliabilities are often lower than validity coefficients (for reasons that we cannot take up here).

Of basic concern to us is the question: Does this measure have any utility for the measurement of motivation in educational settings and specifically in culturally disadvantaged areas? Most of the early research results were disappointing in this regard. Apparently, n Achievement scores are not very highly correlated with school grades or standardized achievement test scores in high-school and college students. As a result, educational interest in the measure lagged after an early flirtation, but it has recently been revived by the evidence that it is possible to change the motive by training. Before we get dazzled by the thought of changing motives in education, we should try to understand the relationship between n Achievement and school behavior. I think there are good reasons to believe that the measure has been misapplied but that it, or a derivative of it measuring the O-P variable, may be very valuable for education.

This misapplication of the n Achievement measure stems from the two most basic levels possible: a lack of understanding of the theoretical basis for the measure, and lack of clarity about its scaling properties.

The problem of understanding the theoretical basis for the n Achievement measure is formidable. Exactly why there should be a relationship between the content of stories written by a subject in a one-half hour period and many other manifestations of his behavior, particularly school behavior, is still a problem shrouded in mystery. I will spare you a discussion of several explanatory attempts that have been made and present one based on the notion that the measure gives a sample of thoughts of the subject.

Let us be very simple-minded and conceive of thoughts as a type of response, arbitrarily remove them from the ghostlike realm of the mind, and relegate them to the more prosaic realm of the physical world of behavioral events. After all, the n Achievement measure is actually based on observable responses, not on pure thoughts. From this point of view, thought samples are a measure of one domain of responses that we wish to correlate with other domains. Notice in passing that we have made no assumptions about a causal chain, that thoughts are antecedent to or "cause" behavior. We have a habit of thinking that thoughts precede actions but there is very little evidence to substantiate it, and in the realm of achievement motivation, there is some evidence that the behavior may precede the thoughts (McClelland, 1966).

The definition of what constitutes an achievement thought is critical in predicting what behaviors will be associated with such thoughts. The scoring manual (Atkinson, 1958) identifies achievement thoughts by the term Achievement Imagery and says they can be seen in a story when one of the characters shows concern over competition with a standard of excellence. The character can show competition when any concrete goal is at stake, such as winning a race, getting a good grade, becoming a success in a career, building a better mousetrap, and so on. The non-specificity of the goal of achievement behavior leads us to the scaling problem mentioned earlier to which we shall return, but let us pursue a little further the theoretical basis for the measure.

Thought-sampling is based on an assumed correlation between thoughts and more concrete behavior.

The advantage of using a sample of thoughts rather than some other type of behavior is this: of all behaviors, thoughts are least subject to inhibition. If we seek a measure of what a person wants to do, we need to rid it of restrictions as to what he can do. For this and other reasons, thought-samples stand a chance of giving us a better picture of a person's generalized motives.

In attempting to measure what a person wants to do, we must follow quite a different road from that taken in measuring what a person is able to do. Put in another way, the measurement of motivation is quite different from the measurement of intelligence, ability, or even attitudes. We want to know what a person will do, not what he can do, although the latter may form a limit in any particular circumstance. In measuring motivation, we sample his thoughts under neutral conditions: not when he is pressured and a pawn, but when he can originate some of his own behavior. By definition, in such a situation, we cannot have control of the kinds of responses observed. If we tighten control, we lose the ability to measure what he will do on his own and end up with a measure of what he can do when we pressure him.

This paradoxical measurement situation is more or less unique and doesn't fit very nicely into some of the most basic conceptions of scaling. To make this a little clearer, think of Skinner's distinction between operants and respondents. Roughly, a respondent is a response that we can elicit directly by manipulation of a specific stimulus. We more or less directly produce a response in the organism by external manipulation. Salivary, pupillary, and eyelid responses are examples. An operant is quite different. Although Skinner assumes that this type of response is probably elicited by some stimulus, we don't know and may never be able to know what it is, so we cannot control the response through the eliciting stimulus. Ignoring control by reinforcement for the time being, we measure an operant by observing its spontaneous emission under conditions of relative freedom.

Let us now conceive of thought sampling as measuring the rates of spontaneously emitted thoughts, i. e., as the measurement of operants. Here we launch into new territory. While Skinner is primarily concerned with shaping a specific operant and converting uncontrolled behavior into behavior that is under the control of the experimenter, we are concerned with measuring a generalized class of operants in the realm of thoughts and relating this measure to other more concrete behavior patterns.

This leap raises many problems, but the major methodological problem is that our measure of a generalized class of spontaneously emitted thoughts cannot be assumed to result in a conventional scale characterized by transitivity along an intensity dimension. Let me try to illustrate this: Pictures used to elicit thought samples from which n Achievement scores are derived show a variety of characters, for instance, a high-school student, two inventors, some businessmen, etc. In what sense can a person (A) who writes achievement stories about inventors, doctors, and businessmen and attains a high score be said to be higher on n Achievement than another person (B) who writes an achievement story about a high-school boy but not about inventors, doctors, and businessmen and receives a relatively low score? If we observe many types of behavior in many situations, we may assume that person A will show more general achievement behavior than person B. But if we measure only school behavior, the reverse may be true. In an unpublished study, we have demonstrated that a six picture measure of n Achievement derived from a standard set does not meet the criterion set by the Guttman technique for unidimensional scalability (Stouffer, Guttman, Suchman, Lazarfeld, Starr, and Clausen, 1947).

Does this mean that the measure is simply unreliable? I think not, although it does mean that it is not a unidimensional measure of the intensity of a motive to do a very specific thing and, therefore, should not be expected to relate to or predict rather specific behavior, such as that expected on a standardized achievement test taken under conditions that produce respondent, or at least discriminated operant, rather than spontaneously emitted operant type behavior. It would be nice to be able to measure motivation along an intensity scale, and most of motivation theory influences us to think of motivation in terms of intensity concepts such as need and drive. If we think back for a moment to the discussion of the Origin-Pawn dimension, however, we can see that a generalized measure of what a person will do when not forced may come closer to our concept of motivation than an intensity measure of what a person can do when pressured.

In the book presenting the original development of the n Achievement measure, McClelland, Atkinson, Clark, and Lowell (1953) briefly mentioned that the technique measured the extensity of Achievement thoughts in a broad realm rather than the intensity of such thoughts. This mention has apparently been lost in the flood of research, most of which was not careful to note the the distinction.

As suggested above, a careful analysis of the basic process involved in the measurement technique gives some insight into the low correlations between n Achievement and academic performance. It gives us more. I think it gives us the basis for suggesting that n Achievement scores will be related to much more general response patterns. Since the measure taps thoughts about achievement in several areas, if a person who has many such thoughts is observed over a period of time in which he may be confronted with some of these situations, we should be able to predict that he will react in achievement ways more often than not. In view of the analysis, it may be expected that the measure of n Achievement will be more valuable in relation to cultural orientations (McClelland, 1961) and long term career patterns of individuals, than in predicting specific responses such as school achievement or exam-taking behavior.

One example may be cited. McClelland (1965) reports that of the original sample of subjects tested at Wesleyan in the validation procedures for n Achievement fifteen years ago, large numbers have gravitated toward entrepreneurial business occupations subsequent to graduating from college. A check of the alumni directory indicated that 83 percent of the men who could be classified as entrepreneurs now, had high n Achievement scores as sophomores in college. Only 21 percent of those classified as non-entrepreneurs had high n Achievement. These results indicate that n Achievement predicts career patterns over a crucial period of life when career choices are being made.

When it comes to measuring motives in culturally disadvantaged children, we have a lot to learn. The data that I have to report results from a study that was initially intended to answer some of the most rudimentary practical questions such as "Can thought samples be collected from this population, i. e., can they produce scorable stories under standardized conditions?" If it turns out that the thought samples are scorable, we can ask whether the n Achievement score is valid for this population. Is the score, for instance, related to risk-taking strategies in this population as it is in middle-class high-school and college students? Since we were interested in beginning to probe the Origin-Pawn dimension in this population, we sought an existant measure that would give an estimate of the children's relative feelings of powerlessness. In short, we tested the feasibility of collecting thought-samples and found fifth, sixth, and seventh grade disadvantaged children quite capable of producing scorable stories. We devised a risk-taking situation that had the bonus of providing us with a measure of school skills, and we used a measure of powerlessness.

The subjects of the study were ninety-four Negro children in two fifth and seventh grade classrooms in one large city (sixty girls and thirty-four boys), and 120 Negro children from several sixth grade classrooms in another large city. Both samples were taken from school areas known to be the most culturally disadvantaged in each city. However, since we were primarily interested in testing the feasibility and validity of the motive measure, we did not sample meticulously, nor did we carefully assess the comparability of the areas. For this reason, comparisons between the two samples are dangerous.

THE MEASUREMENT OF ACHIEVEMENT MOTIVATION

After introductory remarks, the children were handed a six-page form with a different cue typed at the top of each page, and four standard questions equally spaced on each page to help them complete their stories. The verbal cue was read aloud (e. g., "Two men are working at a machine."), and the children were given four minutes to write their stories. In all, six stories were required.

These stories were content analyzed by the procedures described by McClelland, et al. (1953) by two independent scorers. Interscorer reliability was approximately 90 percent agreement between scorers.

THE MEASURE OF POWERLESSNESS

The children's form of the measure of powerlessness (Bialer, 1961) was administered immediately after the measure of achievement motivation. Twenty-three statements are presented in simple language and the child circles either "Yes" or "No" indicating whether he agrees or disagrees with the statement. For example, one item was, "Does it ever help any to think what you will be when you grow up?"

RISK-TAKING BEHAVIOR

In a game situation in which the person can choose to shoot for a fixed target from varying distances of increasing difficulty with increasing reward attendant on success, a group of high-achievement persons will tend to take more shots in the middle range, avoiding both easy and extremely hard shots. A similar group of low-achievement persons will take more shots in the easy and extremely hard positions (cf., Atkinson, Bastian, Earl, and Litwin, 1960). The range of risk which may be considered a moderate risk is dependent upon the skill of the individual and is evidently chosen in terms of where he thinks he has a moderate chance of success (cf., deCharms and Davé, 1965).

We rejected the simple game situation most often used. Instead, it was decided to attempt to develop a new situation completely analogous to the game utilized in our earlier work but involving the typical school behavior of children, i. e. spelling and arithmetic skills.

SPELLING AND ARITHMETIC SKILLS

Each of these measures was comprised of sixty items divided into six levels of difficulty of ten items each. Spelling words were taken from standard lists calibrated for difficulty, and arithmetic problems were taken from a series of arithmetic textbooks used in the schools from early elementary-grade level to junior-high-school level. Extensive pretesting was used to assess empirically the validity of the various levels, and to assure a high degree of success for all children in the easiest level, and increasing difficulty up to a level where little success could be expected from any child.

These measures were administered by standard classroom procedures. The items were arranged in blocks of six, each block progressing from the easiest to hardest level of difficulty in order, though this was not made obvious to the children. It was stressed that this was not a test in the sense that the children would receive grades, but that they themselves would learn from this about their own skill.

After all children had attempted all sixty items, the various levels of difficulty were explained to them and their own ability at each level was impressed upon them by pointing out how many they had completed correctly at each level.

SPELLING AND ARITHMETIC RISK-TAKING TASK

After the group session, the experimenter met with each child individually. She made sure that he understood the various levels of difficulty in the sixty items he had attempted, and how many he had completed correctly at each level.

The child was told that he would now take a test in which he could choose each time the level of difficulty of the item which he would attempt. It was made clear that the test items were divided into levels comparable to those of the previous skill task.

Before he was presented with each item, he was to choose the level of difficulty which he would like to attempt on that trial. The object was to attain "points," the significance of which was left unexplained, as in a game. The points that could be won by correctly completing the item increased with the level of difficulty of the item chosen. These points were established for each child separately, according to the empirically established level of difficulty for him. Thus, a skillful child who chose level five (a difficult level), and who had done four out of ten items correctly at that level in the skill-task (empirical probability of success equals .40), stood to gain thirty-eight points if correct. A less skillful child choosing the same level five may have done only one out of ten items correctly at that level in the skill-task (empirical probability of success equals .10). Such a child could win 150 points if correct at level five. (This procedure, developed by deCharms and Davé (1965), equates subjects as to skill and emphasizes choices in terms of the child's own chances of success.)

Each time the child chose a level for the next item (of ten spelling and ten arithmetic items), he was asked to consult his own paper from the skill-task and report how many he could expect to get right out of ten (his probability of success at that level). He was then asked to consult the following table and report how many points he could win. The point scale was constructed so that the number correct multiplied by the number of points always approximated a constant (150) as closely as possible without using decimals. The expected value of ten trials at any level, therefore, equals 150.

Number you got right out of ten on skill-task	10	9	8	7	6	5	4	3	2	1	0
Points you get if right	15	17	19	22	25	30	38	50	75	150	300

By such means the experimenter made every effort to insure that the child thought of his choices in terms of his own performance on the skill-task and in terms of the possible points which were assigned in accordance with his own demonstrated skill.

Each child made ten choices and attempted ten spelling words, recording his accrued points after each trial. Following this, each child did the same with arithemtic items.

RESULTS

Table 1 presents the mean skill scores on arithmetic and spelling tasks by sex and level of achievement motivation. You will note that the means for the high n Achievement subjects are higher in all cases than the means for low n Achievement subjects. The differences tested by analysis of variance are significant in every case except for arithmetic for girls. There is clearly a relationship overall between n Achievement level and school skills as measured by our test. The relationship is stronger for boys than for girls, but all differences are small, a fact that is attributable to the low variances of the scores. The 60-item task was designed so that the majority of the subjects would success on almost all of the items at the easiest level and fail on almost all at the most difficult level. As a result, the range of probable scores is much smaller than the total range of sixty, and the variances are quite small. In addition, these results come from a larger analysis where variance attributable to other factors, such as grade level, was extracted.

Table 2 presents data from the same dependent variables, spelling and arithmetic skills, but this time showing the relationship with the I-E scale. The low I-E subjects feel relatively powerless. The relationship is significant in both cases of arithmetic and spelling skills for girls but not for boys. This sex difference is more or less the opposite of what was found with n Achievement. These results seem to indicate that the thought sample measure is more powerful in predicting school skills for boys, whereas the questionnaire measure is more powerful for girls. The data by themselves are not very impressive in this respect, and might seem merely a function of the specific tests except for two things. First, we have argued that thought sampling results in a different kind of score from that derived from a questionnaire, and the difference is interesting theoretically. Second, a recent review of the literature of n Achievement indicates a similar trend toward an interaction between sex and type of testing situation. Klinger (1966) reviewed some forty-four studies that investigated the relationship between measures of achievement motivation and task performance. Some of the studies used the thought-sample technique which Klinger calls the TAT, and some of them used a measure called the Iowa Picture Interpretation Test, or the IPIT. This test uses pictures but supplies several plots for the subject to choose from in a multiple choice form. Klinger says: "The studies also suggest an interaction of Instrument × Sex of Subject, in that of the studies that used the TAT, a higher proportion of those with male subjects reported significance than those with females, while the reverse was true with the IPIT" (Klinger, 1966, pp. 297-298). Klinger makes no attempt to explain this interaction.

In view of our analysis of thought sampling as a measure of a more or less "free" operant where the subject emits responses, whereas a questionniare or multiple choice measure supplies ready-made responses for the subject to choose from, it seems reasonable to hypothesize that the free operant measure may be more appropriate for male subjects and a respondent (or discriminated operant) measure more appropriate for females. I cannot pretend to explain in detail why this should be so, but it does seem plausible, especially in the achievement area, and in our culture, that men are expected to produce achievement behavior spontaneously, whereas women may have to respond appropriately with achievement behavior when pressed but are not necessarily expected to initiate it on their own. We hope to investigate this possibility in the near future.

The results from the risk-taking data are presented in Tables 3 and 4, and may be seen in Figures 1 and 2. The means in the tables need a little explanation. The hypothesis derived from earlier studies predicts that subjects with high achievement motivation will more often choose spelling words or

TABLE 1

MEAN SPELLING AND ARITHMETIC SCORES (OF A POSSIBLE 60) BY SEX AND LEVEL OF *n* ACHIEVEMENT

		Boys			Girls		
		Math**	Spelling**	Total	Math	Spelling*	Total
High *n* Achievement	N	46	46		60	60	
	$\bar{x}$	21.7	28.2	24.9	22.5	31.0	26.8
Low *n* Achievement	N	51	51		63	63	
	$\bar{x}$	18.5	21.7	20.1	19.6	26.5	23.1
Total		20.0	24.8	22.4	21.0	28.7	24.9

* *p*. $<.05$
** *p*. $<.01$

TABLE 2

MEAN SPELLING AND ARITHMETIC SCORES (OF A POSSIBLE 60) BY SEX AND I-E LEVEL (POWERLESSNESS)

		Boys			Girls		
		Math	Spelling	Total	Math*	Spelling*	Total
High I-E	N	46	46		69	69	
	$\bar{x}$	20.9	25.2	23.1	22.6	29.8	26.2
Low I-E	N	51	51		54	54	
	$\bar{x}$	19.3	24.4	21.9	19.0	27.2	23.1
Total		20.0	24.8	22.4	21.0	28.7	24.9

**p*. $<.05$

arithmetic problems in the middle ranges of probability levels from say 30 percent to 70 percent probability, while subjects with low *n* Achievement will not show this tendency to as great an extent and may, in fact, avoid the middle ranges and choose more extremes either around 90 percent or 10 percent.

In order to extract a measure from the data to test this hypothesis statistically, we used a procedure developed by deCharms and Davé (1965). The average deviation of the ten probability choices for each subject is measured around the overall mean of all choices for all subjects. A large average deviation indicates that the individual subject's choices range farther out in either direction on the scale; a small average deviation indicates that the subject's choices are concentrated in the middle range. The hypothesis then predicts that subjects with high *n* Achievement will have smaller average deviation scores than subjects with low *n* Achievement. Tables 3 and 4 present the means of the average deviation scores from each group. We shall call this score the risk score.

Table 3 gives strong support for the hypothesized relationship between moderate risk-taking and high *n* Achievement under all conditions, i.e., for both boys and girls and when the task involves either spelling or arithmetic problems. Table 4 indicates that the I-E scale is not related to risk-taking under any conditions.

A look at Figures 1 and 2 may make the results with *n* Achievement clearer. Figure 1 combines

TABLE 3

MEAN RISK-TAKING SCORES BY SEX AND LEVEL OF *n* ACHIEVEMENT
(LOW MEAN INDICATES MODERATE RISK)

		Boys			Girls		
		Math*	Spelling***	Total	Math*	Spelling***	Total
High *n* Achievement	N	46	46		60	60	
	$\bar{x}$	19. 2	24. 2	21. 7	22. 0	23. 3	22. 6
Low *n* Achievement	N	51	51		63	63	
	$\bar{x}$	23. 8	29. 1	26. 4	24. 4	27. 9	26. 1
Total		21. 6	26. 8	24. 2	23. 2	25. 6	24. 4

* $p. < .05$
***$p. < .005$

TABLE 4

MEAN RISK-TAKING SCORES BY SEX AND LEVEL I-E (POWERLESSNESS)
(LOW MEAN INDICATES MODERATE RISK) *

		Boys			Girls		
		Math	Spelling	Total	Math	Spelling	Total
High I-E	N	46	46		69	69	
	$\bar{x}$	21. 7	27. 5	24. 6	22. 7	25. 3	24. 0
Low I-E	N	51	51		54	54	
	$\bar{x}$	21. 5	26. 2	24. 4	23. 8	26. 0	24. 9
Total		21. 6	26. 8	24. 2	23. 2	25. 6	24. 4

*No significant differences

boys and girls and shows a high peak in the mid ranges of probabilities of success for the high *n* Achievement subjects on arithmetic problems. The curve for the low *n* Achievement subjects shows an interesting propensity to take extremely speculative risks. For the low *n* Achievement subjects 58 percent of the choices are in the range of 0 to 40 percent probability of success.

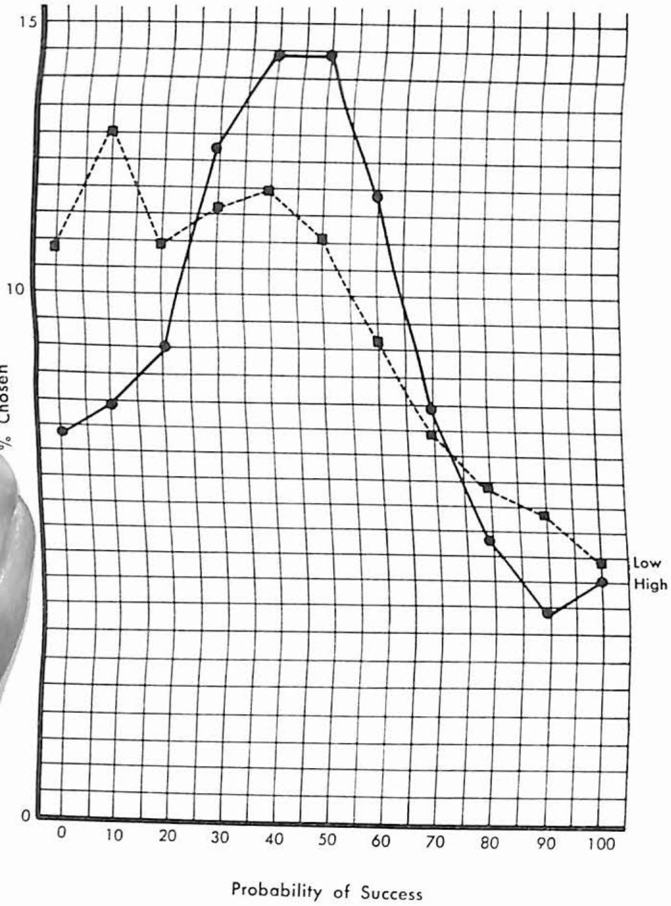
Figure 2 presents the results for spelling. Here the high *n* Achievement curve shows a hump in the middle, and the low *n* Achievement curve actually shows a trough in the middle ranges and is higher at both extremes as predicted by the hypothesis.

These results with arithmetic and spelling tasks show a clearer differentiation in risk-taking strategies between high and low *n* Achievement subjects than most previous studies in this area, and suggest the possibility that risk-taking on a serious task such as spelling or arithmetic is more clearly a function of *n* Achievement than on the game-like tasks used in other studies. It is also possible that the effect of achievement motivation may be more clearly evident in the culturally disadvantaged environment than in the highly competitive middle class culture where the range of *n* Achievement scores may be skewed toward high *n* Achievement.

DISCUSSION

All of the results reported from this study were derived from two different samples that were combined for simplicity of presentation. When the samples are investigated separately, the findings reported replicate in almost every case.

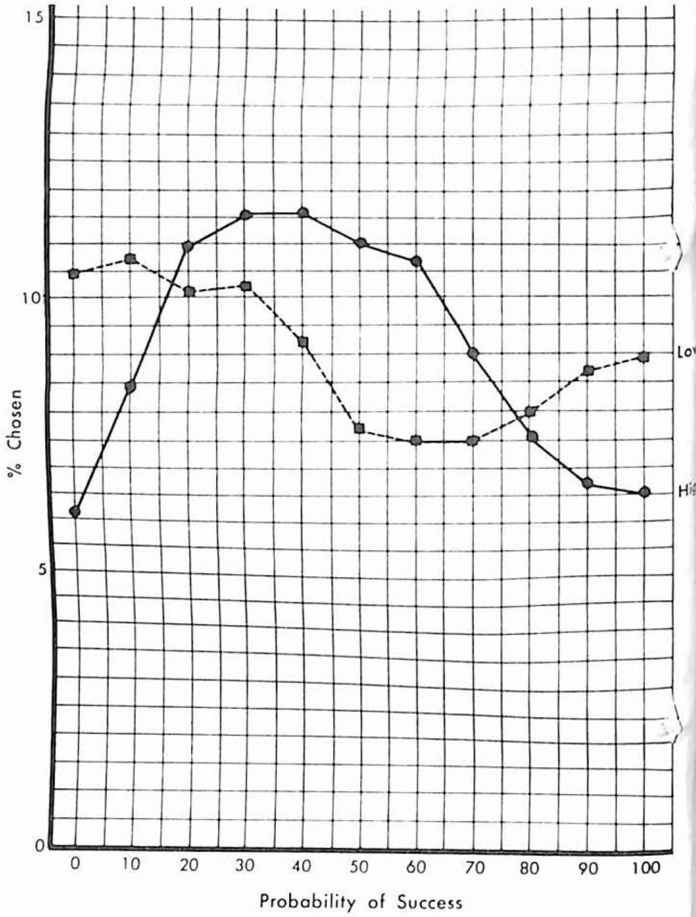
Figure 1. Percentage of tasks chosen at each level of probability of success on the arithmetic risk-taking task by subjects above and below the median n Achievement score.



Replication of the data on spelling and arithmetic skills is especially important since the tests from which these scales derived were intended originally merely to give preliminary data for the risk-taking task. The fact that the relationships with n Achievement are significant at all is surprising and even embarrassing since the major thrust of our theoretical argument predicts that the n Achievement measure should be more highly related to measures of change over time than to single test measures such as these arithmetic and spelling skill tasks. Actually, the results, though significant, are not indicative of high correlations, but of relationships quite commensurate with others reported between n Achievement and academic achievement. Although we have argued that n Achievement is primarily a measure of the extensity of thoughts about achievement and should, therefore, be related to long-term trends in behavior, our data remind us that a measure of extensity is not completely independent of intensity. To use the extensity measure exclusively as a measure of intensity is to misapply it; but the extensity of thoughts about achievement may give some indication of the intensity of achievement motivation is a specific area.

Probably the most interesting result of these preliminary forays into the study of motivation in culturally disadvantaged children is the evidence that the thought sample measures are more effective with boys than girls, and with the questionnaire measure the reverse is true. If this result stands up on replication, it may lead to further insights about thought sampling.

Figure 2. Percentage of tasks chosen at each level of probability of success on the spelling rish-taking task by subjects above and below the median n Achievement score.



It should be clear by now that the I-E scale used here was not designed to measure the origin-pawn dimension. It is our hope that a measure of this dimension can be devised based on thought sampling. To be valid, the measure must allow the subject to emit his own behavior. Choice of reactions selected and presented by the experimenter as on a questionnaire limits the subject. We are at present beginning to develop a measure based on thought samples, and preliminary evidence indicates that it is feasible and may be related to the measures of school skills discussed here and even to long-term indications that a person prefers to take personal responsibility for his actions.

We can only end this paper by noting that it is only a beginning. The research reported here constitutes the basis for a project that is just getting under way, the aims of which are to relate n Achievement to change in academic skills in a longitudinal study, to develop the concept and measure the origin-pawn dimension, and ultimately to develop training techniques for teachers and pupils similar to those used in achievement motivation training courses with businessmen (McClelland, 1965).

REFERENCES

Atkinson, J. W. (Editor) Motives in fantasy, action, and society. Princeton, New Jersey: Van Nostrand, 1958.

Atkinson, J. W., Bastian, J. R., Earl, R. W., and Litwin, G. H. The achievement motive, goal setting, and probability preferences. Journal of Abnormal and Social Psychology, 1960, 60, 27-36.

Bialer, I. Conceptualization of success and failure in mentally retarded and normal children. Journal of Personality, 1961, 29, 303-320.

deCharms, R., Carpenter, Virginia, and Kuperman, A. The "origin-pawn" variable in person perception. Sociometry, 1965, 28, 241-258.

deCharms, R., and Davé, P. N. Hope of success, fear of failure, subjective probability, and risk-taking behavior. Journal of Personality and Social Psychology, 1965, 1, 558-568.

deCharms, R., Dougherty, K., and Wurtz, S. The origin-pawn variable. Unpublished manuscript, Washington University, 1965.

Heider, F. The psychology of interpersonal relations. New York: Wiley, 1958.

Hull, C. L. Principles of behavior: An introduction to behavior theory. New York: Appleton-Century, 1943.

Klinger, E. Fantasy need achievement as a motivational construct. Psychological Bulletin, 1966, 66, 291-308.

Kuperman, A. Relations between differential constraints, affect, and the "origin-pawn" variable. Unpublished doctoral dissertation, Washington University, 1966.

McClelland, D. C. Personality. New York: Sloane, 1951.

McClelland, D. C. The achieving society. Princeton, New Jersey: Van Nostrand, 1961.

McClelland, D. C. Toward a theory of motive acquisition. American Psychologist, 1965, 20, 321-333.

McClelland, D. C., Atkinson, J. W., Clark, R. A., and Lowell, E. L. The achievement motive. New York: Appleton-Century-Crofts, 1953.

Mowrer, O. H. Learning theory and behavior. New York: Wiley, 1960.

Pepitone, A. Attributions of causality, social attitudes, and cognitive matching processes. In R. Tagiuri and L. Petrullo (Editors), Person perception and interpersonal behavior. Stanford: Stanford University Press, 1958.

Rotter, J. B., Seeman, M., and Liverant, S. Internal versus external control of reinforcements: A major variable in behavior theory. In N. F. Washburne (Editor), Decisions, values and groups. Vol. 2. New York: Macmillan, 1962.

Ryle, G. The concept of mind. New York: Barnes and Noble, 1949.

Smith, L. M., and Geoffrey, W. Toward a model of teacher decision-making in an urban classroom. Cooperative Research Branch. Project 5-048, 1965.

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarfeld, P. F., Starr, Shirley A., and Clausen, J. A. The American soldier: Measurement and prediction. Princeton, New Jersey: Princeton University Press, 1949.

Thibaut, J. W., and Riecken, H. W. Some determinants and consequences of the perception of social causality. Journal of Personality, 1955, 24, 113-133.

Young, P. T. Motivation and emotion: A survey of the determinants of human and animal activity. New York: Wiley, 1961.

PART II

Subject Matter, Content and Sequence

CURRICULUM improvement in the various disciplines has been hampered by lack of adequate research paradigms to determine the effectiveness of various instructional programs, by lack of relevant learning theory, and by inadequate conceptualizations for determining the content to be learned. The structure of the discipline itself, as formulated by one or more scholars, has provided much of the basis for recent curriculum development. E. G. Begle in "Curriculum Research in Mathematics" cites the extensive work of the School Mathematics Study Group in the development and evaluation of a mathematics curriculum. The structure of mathematics has proven to be a powerful organizer for the mathematics curriculum at the university, high school, junior high school, and now at the elementary school levels. The point of view adopted is that in learning mathematics each pupil is to build his own structure of mathematics. Such a structure starts with a small number of concepts and eventually incorporates strategies of problem solving. Begle argues that the necessary framework would be characterized by three aspects: "1) the mathematical content of the curriculum, 2) a theory of learning of the content, and 3) a theory of the application of the content to problem solving." Begle reports that some of the results obtained by the School Mathematics Study Group are being investigated in a National Longitudinal Study of Mathematical Abilities.

Similar problems of curriculum research and development are common to science. Arthur Livermore examines the structure of science (a science of science) as the basis for curriculum reform. Reports of the various curriculum groups (PSSC in physics, BSCS in biology, and CHEM and CBA in chemistry) indicate a union of scientists, educators, and psychologists working together on the problem of curriculum design and evaluation. The evolution of the so-called new curriculum in science can be characterized as starting effectively first at the high school level, then junior high school, and now at the elementary school level. While these curriculum studies emphasize concepts of importance within the respective disciplines, corresponding emphasis is also given to the empirical aspects of science as represented in the laboratory activities common to all of the curriculum studies. For example, the American Association for the Advancement of Science has produced a program for the elementary school which emphasizes the process approach to science. In this approach certain cognitive skills are used as the organizers of the curriculum. However, the lack of an encompassing conceptual framework in the science inhibits the development of a systematic curriculum K-12. Rather several well-prepared curricula are being developed at the various school levels.

Curriculum development in English is also dependent on the structure of that discipline. Robert Pooley discusses various efforts at developing curriculum rationales and materials in such areas as literature, composition, language, and in the emerging grammars of the English language. Development and research underway in the various areas are reported with particular emphasis given to the structural and transformational grammars.

Reading ability is of fundamental concern in the usual curriculum. The demands placed on the learner vary across age, and across subject matter, yet the student is not trained in techniques of reading science, mathematics or English. A research paradigm appropriate to investigating variation in the rate of reading is reported by Wayne Otto, Theodore Harris, and Thomas Barrett in "Research in Reading." The authors investigate the question of "whether variability in reading speed developed by training with carefully constructed, short model paragraphs will transfer to reading the less tightly structured passages typically found in classroom material." Extensive details on procedures and results are reported.

The contribution of psychological research to the development of curricula in mathematics is expanding, according to Myron Rosskopf. He cites the work of various psychologists in discussing the attainment of mathematics concepts. The effect of the learner's system of focusing on the problem at hand and the available strategy of problem solving are discussed.

Curriculum Research in Mathematics

E. G. Begle
Stanford University
Stanford, California

I SHALL interpret "curriculum" in the above title very broadly, since I believe that research in mathematics education must involve research on the learning of mathematics, the teaching of mathematics, problem solving, etc., as well as the syllabus, and that exclusive concentration on any one aspect of this complex is not fruitful.

I shall not attempt a historical survey of research in mathematics education. Weaver (1957-1966) has provided useful lists covering research on elementary-school mathematics for each year since 1951. Brown (1958, 1960, 1963, 1965) has provided biennial annoted reports of research on mathematics teaching carried out since 1956. A glance at these reports indicates that a considerable amount of research has been done in the past and also that the volume of research is increasing rapidly. A more careful look, however, provides the melancholy information that this large number of research efforts yields very little that can be used to improve mathematics education. This is not to say that all these efforts have been wasted. There are in fact many ingenious ideas and suggestions which ought to be followed up. In general, though, the results are too special or incomplete to be of any wide use.

It does not take much reflection to see why this should be the case. Research in education in general, and in mathematics education in particular, is much more difficult than research in say, psychology or chemistry. For one thing, the number of variables that should be taken into account is very large. Mathematics achievement can be affected, at least in theory, by the prior learnings of the subjects, by a wide variety of cognitive abilities and cognitive styles, by attitudes toward mathematics, and, as we well know, by socio-economic and demographic variables. Similarly the cognitive and affective state of the teacher may well affect the student's learning of mathematics.

In order to isolate and compare two different curriculum treatments, for example, it is necessary to take all these variables into account so that the effects due to these variables will not be confounded with those due to the curriculum treatments. To do this generally requires either a large experiment or a large number of carefully correlated small ones.

Until very recently, it has been almost impossible to meet this requirement. Experiments in mathematics education have been carried out by individuals with limited resources. The experiments of necessity were small in scale, involved relatively small numbers of subjects, and had a short time span. Only very recently has financial support become available which makes possible some of the studies and experiments necessary for significant improvement in mathematics education.

A large scale investigation now being carried out by the School Mathematics Study Group is the National Longitudinal Study of Mathematical Abilities (Cahen, 1965).

In this study three groups of students entering grades 4, 7, and 10 in September, 1962, totaling over a hundred thousand, will have been followed for five years. The mathematics achievement of each of these students has been measured twice each year.

The primary purpose of this study is to ascertain the effects of the new mathematics curricula in order to provide guidance for further curriculum development.

It was necessary, of course, to measure not only the level of mathematics achievement of each student at the beginning of the study, but also as many as possible of the other variables which might affect the results. We have therefore obtained measures of a considerable number of cognitive variables, selected by our psychological consultants, for their probable relevance to mathematics achievement. An extensive attitude inventory was administered three times during the study, since we want to know how attitudes affect achievement and also how achievement affects attitudes. We have some information about the socio-economic status of the students as well as information about the schools. Most of the teachers of the students included in the study have provided us with extensive information about their training and experience, and more than half of them have also provided information about their attitudes toward mathematics, teaching, etc.

When we analyze this enormous mass of information, we will not only find out what we want to know about the differential effects on mathematics achievement of modern curricula as compared with conventional ones, but we will also obtain as a by-product, information on the effects on mathematics achievement of the other variables we have measured.

This kind of information should prove valuable both to the design of future experiments and in the interpretation of past results. Hopefully we will find that many of these variables have little or no effect on mathematics achievement. When this is the case, future experiments involving similar students need not take such a variable into account, and we will have more confidence in generalizing the results of past studies which did not control on such a variable.

However, while studies such as this one will prove very helpful, it is doubtful that they will prove sufficient. There is another fundamental problem which must be overcome before we can expect significant improvements in mathematics education. This problem is the lack of any theoretical framework for mathematics education. Without such a framework, one whose broad outline has reasonably widespread acceptance, there is no way to program the many individual small studies which will continue to be undertaken so that they can be mutually reinforcing and can contribute to overall forward progress. It is my belief that the time is ripe to begin the construction of such a theoretical framework.

THE MATHEMATICAL CONTENT OF THE CURRICULUM

One aspect of it of course will be concerned with the mathematical content of the curriculum. Before discussing this, I will turn to what may seem at first to be a digression. I am sure that many of our teachers look on our new curricula as a revolution in school mathematics. In a sense this is correct, but it is merely a small aspect of a revolution in mathematics itself which has been going on for a century and a half.

This revolution stems from the work of Abel and Galois. When Abel demonstrated the nonexistence of an algebraic formula for the solution of the quintic equation and when Galois applied his methods to settle the general problem, the direction of the development of mathematics was changed forever. Obviously they did not obtain their results by time-honored computational procedures. Instead they demonstrated that an examination of the over-all structure of a mathematical system, in contrast to computations with the individual elements of the system, was a very powerful mathematical tool and could lead to solutions of problems not otherwise solvable.

Although the revolution dates back to the previous century, it gathered momentum slowly, and it was only in the present century that its effects became overwhelming. The tremendous flowering of modern algebra and of functional analysis which started early in this century could not have taken place without the ideas about structure first pointed out by Abel and Galois.

During the 1930's these ideas began to be applied in the graduate curriculum, and it became clear that these ideas, already demonstrated to be powerful and indispensable in mathematical research, were also powerful and effective in mathematics instruction at this level. During the next two decades a start was made on a revision of the undergraduate curriculum in the universities along the same lines.

Consequently, when mathematicians joined with high-school teachers in an attempt to improve the secondary-school curriculum, their first inclination was to apply to the subject matter normally taught in these schools the same point of view toward mathematics, an emphasis on structure, which had proved successful not only in mathematical research but also in mathematical education at the university level. When it was discovered that this could be done successfully for the secondary-school curriculum, the reform movement continued and revised the elementary-school curriculum in the same spirit.

There are two important observations which now need to be made. The first is that this revolution has been successfully concluded. The second is that no new revolution is clearly in sight. Even if the first stirrings of a new revolution might be taking place in mathematical research, its effects could not appear in the pre-college program for generations. Consequently we can agree on the broad outline of the content of the mathematics curriculum for the schools.

This content is well enough known that we need not spell it out here in detail. Suffice it to mention that the mathematical topics included in the curriculum are, with few exceptions, those included in the pre-revolutionary curriculum. Only a few new topics, such as inequalities, have been added because of their intrinsic importance, and only a few other topics have been given lesser emphasis than before.

On the other hand, the new curriculum differs radically from the old in that it includes the structure of the common mathematical systems, the basic mathematical concepts and their interrelationships, as well as the basic mathematical facts and techniques.

Before discussing the implications of this new aspect of the curriculum for our theoretical framework for mathematics education, it is well to point out that while few new topics have been incorporated in the curriculum so far, there will undoubtedly be new ones to be added from time to time. The nature of our society at present is such that it would be wise to include as soon as possible, and for all students, some of the basic ideas of probability and statistics as well as some of the mathematical ideas relevant to high speed computers. Incorporation of any such new topic into the curriculum will require research to determine the prior understandings needed by the student, the most efficient location of the topic in the sequence, etc. We can certainly look forward to a continuing need for this kind of research.

Returning now to the curriculum, we see that our aim in mathematics education is to help each student to build in his mind a conceptual structure. This structure, for primary-school children, starts with a small number of concepts, abstracted directly from the real world, about number and geometric configurations. As the child progresses through school, these concepts are refined and sharpened, and the relationships between them became increasingly complex. At some stage, new concepts are met which come not from the real world but rather from previously learned concepts. The concept of irrational number is a case at point. Among these concepts there are, of course, those concerned with computation. For example, we do not ask our students to memorize an infinite multiplication table. Rather, we let them stop after 9×9 and teach them the algorithm for multiplication, and this is a concept.

Naturally, the order in which mathematical concepts are learned is subject to some variation, and much research time could be devoted to studying the effects of variations in the sequence. However, in my opinion, most of this time would be more usefully employed if devoted to some of the problems discussed below.

A THEORY OF THE LEARNING OF THE CONTENT

What this means for us is that our theoretical framework for mathematics education must have as one of its many aspects a theory of the learning of mathematical concepts and relations between these concepts.

No well-developed theory of this sort now exists. One's first thought, when the topic of learning of any sort is brought up, is to turn to our colleagues in psychology for advice and assistance. A very recent survey of concept growth in children (Wallace, 1965) shows that psychologists can provide very few useful answers to our questions. Much work has been done in this country on the development of concepts, but interest is usually focused on the strategies used by the subjects in trying to guess which combinations of color, sizes, shapes, etc., the experimenter has chosen. Neither the concepts nor the strategies have much relevance to classroom learning.

An interesting by-product, however, of these studies is that the availability of verbal labels for a concept plays an important role (Kuenne, 1946) and will have to be taken into account in any theory of the learning of mathematical concepts.

Some of the investigations of "discovery" teaching in this country have involved the learning of some non-trivial mathematical concepts. Unfortunately, the results of these studies present a confused picture. For a review of the present state of affairs, see Cronbach (1966).

Outside the U. S., investigations of concept learning have been dominated by the work of Piaget. He deserves great credit, of course, for investigating real mathematical concepts and how children come to acquire them. Unfortunately, he generally confines himself to the initial concepts of any branch of mathematics and has little to tell us about the learning of the later and more refined concepts or the learning of relationships between concepts.

A more fundamental difficulty in applying Piaget's findings in mathematics education lies in the fact that, when he questions a child to see how well he understands a concept, he is very careful never to provide the child with any subsidiary concepts which might help him. In fact, he seems to be investigating what children can learn without the benefit of education. There are, however, some indications (Kohnstamm, 1965) that the picture is quite different if children are provided with some assistance in the form of subsidiary concepts.

In short, the task of building a theory of the learning of mathematical concepts still lies before us. Some ideas about what will have to be included in such a theory can be obtained from the sources just mentioned. It should also be mentioned that some very interesting work, with which we are just beginning to become acquainted, is being done in Russia, and this work may provide us with further clues.

In order to construct a suitable theory of the learning of mathematical concepts, and such a theory should be quantitative rather than purely qualitative, it will be necessary to construct instruments which measure how well a student understands a particular concept.

This will not be an easy task. Concepts are not observable. It will be necessary to use indirect methods to assess a student's understanding of a particular concept. Perhaps we can obtain some guidance on this from our colleagues in the physical sciences, who also deal with unobservables such as neutrinos.

Perhaps the most useful contribution which Piaget has made to our present efforts is his demonstration that skillful questioning can provide precise and detailed information about an individual's understanding of a concept or problem. Similar individual interrogations will undoubtedly be an important tool in the development of instruments for measuring understanding, although the final instruments will need to appear in a more standardized format. It is worth observing that Brownell (1963) also has found working with individual children a valuable procedure.

THE APPLICATION OF THE CONTENT TO PROBLEM SOLVING

A final aspect of our theoretical framework for mathematics education is that of problem solving. It would be futile to provide our students with a thorough understanding of mathematics if they were not able to use this understanding to solve problems. What we need therefore is a better understanding of how we can help our students to develop useful strategies for solving problems.

This seems to call for a considerable amount of exploratory work. We do have already, however, a variety of problem-solving strategies that can be investigated. Polya (1945) for example, has written extensively on this topic. In another direction, Newell, Shaw, and Simon (1959) have investigated, by means of computer simulation, the power of certain specific strategies. These strategies had been identified as ones actually used by at least some students, and a computer program then verified that the strategies were indeed effective. In still another direction is the very interesting work by Crutchfield and Covington (Crutchfield, 1966). Using comic-book format for their materials, they were able to train their subjects to examine the relationships between the various aspects of a problem situation, and the implications of these relationships, for a sufficiently long time and in sufficient depth to arrive at a solution. Their problems were not mathematical, but their procedures may well have implications for mathematics education.

The Russian literature, mentioned above, also has suggestions of problem-solving strategies which are worth investigating.

There are still two other sources which may provide us with raw material for investigation of problem solving. The strategies used by Newell, Shaw, and Simon were obtained by having a number of students solve problems aloud during individual interviews. This procedure may well prove very fruitful, especially at the elementary school level, where we seem to know less than at the higher levels. Similarly, classroom observations of teachers teaching problem solving and of their techniques for helping students to solve problems may pay good dividends.

In short, we already know a number of problem-solving strategies and we have suggestions for locating others. What we need to do now is to start a systematic study of techniques for teaching these strategies to students and of the ability of students to use these strategies. We need to study in particular the degree of understanding of the conceptual structure of mathematics a student needs before he can learn a particular strategy and use it effectively. This points up again the need for instruments which can be used to measure understanding of mathematical concepts.

To summarize briefly, I have been arguing that substantial improvements in mathematics education will have to be based on more fundamental research than has been carried out so far. I have further argued that the magnitude of the task is such that the efforts of a large number of research workers will be needed and that these efforts can be made mutually reinforcing and contributive to substantial improvement only if there is an overall theoretical framework which is generally accepted and which can be used not only to suggest individual research problems but also to relate individual findings in a meaningful pattern.

Three aspects of such a theoretical framework were mentioned: the mathematical content of the curriculum, a theory of the learning of the content, and a theory of the application of the content to problem solving. The first is, except for minor variations, dictated by the nature of mathematics itself. It differs from the curriculum content of a generation ago by emphasizing the conceptual structure of mathematics as well as routine mathematical skills. This part of our theoretical framework, then, is generally agreed to. The rest however, we will have to construct. We have a few bits and pieces to start with, but a great deal of empirical work will have to be done before even the outlines of these aspects of our general theory will emerge.

Finally, I have argued that an essential tool that we must have to carry out our task is a battery of instruments which measure understanding of mathematical concepts, and hence that our first step should be that of starting to construct such instruments.

CONCLUSION

In conclusion, I should like to note that any framework we may construct for mathematics education may be quite specific and not immediately generalizable to other subject matter areas. In mathematics we deal with only a very few aspects of the real world, namely the numerical and geometric aspects of physical situations. In science however, for example, many more different aspects of reality are considered, e.g., life, energy, mass, acceleration, color, density, galaxies, elasticity, genetics, the periodic table. However, as a student progresses through the curriculum, his mathematical concepts are refined, sharpened, and generalized to a much greater extent than his scientific ones, so the resulting mathematical conceptual structure differs markedly from his scientific one. Hence the process of concept formation in mathematics may not be the same as in science, and problem-solving strategies may have quite different values in the two areas. And when we move to more remote curriculum areas, such as social science or literature, for example, the situation is probably even more different.

REFERENCES

Brown, K. E. Analysis of research in the teaching of mathematics: 1955 and 1956. Bulletin No. 4. Washington, D. D.: U.S. Office of Education, 1958.

Brown, K. E., and Kinsella, J. J. *Analysis of research in the teaching of mathematics: 1957 and 1958.* Bulletin No. 8. Washington, D.C.: U.S. Office of Education, 1960.

Brown, K. E., Kinsella, J. J., Fehr, H., Myers, S., and Weaver, F. *Analysis of research in the teaching of mathematics: 1959 and 1960.* Bulletin No. 12. Washington, D. C.: U. S. Office of Education, 1963.

Brown, K. E., and Abell, T. L. *Analysis of research in the teaching of mathematics.* Bulletin No. 28. Washington, D. C.: U.S. Office of Education, 1965.

Brownell, W. A. Arithmetical abstractions – progress toward maturity of concepts under differing programs of instruction. *The Arithmetic Teacher*, 1963, *10*, 322-329.

Cahen, L. S. An interim report on the national longitudinal study of mathematical abilities. *The Mathematics Teacher*, 1965, *58*, 522-526.

Cronbach, L. J. The logic of experiments on discovery. In L. S. Shulman and E. R. Keislar (Editors), *Learning by discovery: A critical appraisal*. Chicago: Rand McNally, 1966. Pp. 76-92.

Crutchfield, R. S. Creative thinking in children: Its teaching and testing. In O. G. Brim, R. S. Crutchfield, and W. H. Holtzman (Editors), *Intelligence: Perspectives 1965.* New York: Harcourt, Brace and World, 1966. Pp. 33-64.

Kohnstamm, G. A. Developmental psychology and the teaching of thought operations. *Paedogogica Europaea*, 1965, *1*, 79-97.

Kuenne, Margaret R. Experimental investigation of the relation of language to transposition behavior in young children. *Journal of Experimental Psychology*, 1946, *36*, 471-490.

Newell, A., Shaw, J. C., and Simon, H. A. *Report on a general problem-solving program.* Paris, France: Proceedings of the International Conference on Information Processing, UNESCO House, 1959.

Polya, G. *How to solve it*. Princeton, New Jersey: Princeton University Press, 1945.

Wallace, J. G. *Concept growth and the education of the child: A survey of research on conceptualization*. Occasional Publication Series No. 12. Slough, U. K.: National Foundation for Educational Research in England and Wales, 1965.

Weaver, J. F. Six years of research on arithmetic instruction: 1951-1956. *The Arithmetic Teacher*, 1957, *4*, 89-99.

Weaver, J. F. Research on arithmetic instruction – 1957. *The Arithmetic Teacher*, 1958, *5*, 109-118.

Weaver, J. F. Research on arithmetic instruction – 1958. *The Arithmetic Teacher*, 1959, *6*, 121-132.

Weaver, J. F. Research on arithmetic instruction – 1959. *The Arithmetic Teacher*, 1960, *7*, 253-265.

Weaver, J. F. Research on elementary-school mathematics – 1960. *The Arithmetic Teacher*, 1961, *8*, 255-260. (a)

Weaver, J. F. Research on elementary-school mathematics – 1960 (concluded). *The Arithmetic Teacher*, 1961, *8*, 301-306. (b)

Weaver, J. F. Research on elementary-school mathematics – 1961. *The Arithmetic Teacher*, 1962, *9*, 287-290.

Weaver, J. F. Research on elementary-school mathematics – 1962. *The Arithmetic Teacher*, 1963, *10*, 297-300.

Weaver, J. F. Research on elementary-school mathematics – 1963. *The Arithmetic Teacher*, 1964, *11*, 273-275.

Weaver, J. F. Addenda to the May, 1965 report on research in elementary school mathematics. *The Arithmetic Teacher*, 1965, 12, 577-578. (a).

Weaver, J. F. Research on elementary-school mathematics – 1964. *The Arithmetic Teacher*, 1965, *12*, 382-387. (b).

Weaver, J. F. Research on mathematics education, grades K-8, for 1965. *The Arithmetic Teacher*, 1966, *13*, 414-427.

Curriculum Research in Science

Arthur H. Livermore
American Association for Advancement of Science
Commission on Science Education
Washington, D. C.

DEVELOPING and testing new science programs for grades K through 12 has been an activity of hundreds of university scientists and school teachers during the past ten years. The widespread and large-scale revolution in pre-college science teaching has been possible because large sums of money have been available from the National Science Foundation to finance the activities. The early curriculum reforms came at the high school level and these are now well established in school systems throughout the country. Because they are well known I will not elaborate on them, but it is important to review them briefly because they have, in large measure, set the pattern for science curriculum research in the pre-high school grades.

HIGH SCHOOL CURRICULUM RESEARCH

These high-school curriculum projects did several important things. In the first place they stripped away the authoritarian discussions of modern technological advances which had been tacked on to science courses that were twenty years or more out of date. They adopted the philosophy that teaching science was much more than teaching a catalog of facts. Rather, it should be teaching facts together with concepts and developing at the same time an understanding of and an ability to engage actively in the scientific enterprise.

To accomplish these things the approach to teaching was radically changed. No longer was the discipline presented in lecture and in text in didactic fashion. No longer did the laboratory serve as a place where the student demonstrated to himself facts already known. Instead, the laboratory has become the place where facts are learned through investigation and the classroom the place where the results of laboratory investigations are discussed and an understanding of fundamental scientific concepts are developed.

The Physical Science Study Committee (PSSC) led off the revolution in 1956. This group decided to deal with physics as a unified story in which the successive topics would lead toward an atomic picture of matter, and along with all this they hoped to develop the idea that physics is not a book completed or closed but is an unfinished and continuing activity. They designed new and simple laboratory equipment and devised new experiments so that the PSSC laboratory became the place where students carried on investigations, not simply demonstrations.

In a similar way, the other high-school curriculum groups pointed their courses in new directions.

The Chemical Bond Approach (CBA) project grew out of a conference of college scientists and high-school teachers held in the summer of 1957. The group noted that little change had taken place in high-school chemistry courses for at least twenty years and that modern chemical concepts were little mentioned and little used. The CBA group decided to make the most fundamental of all chemical concepts the central theme of the program. Hence, the idea of chemical bonds has become the unifying thread of the curriculum. Chemistry is presented as an interplay of experiment and ideas – of what is observed and what is thought. In the laboratory program the students carry out small investigations, and in the classroom they learn to apply chemical concepts to explain their observations.

The Chemical Education Materials Study (CHEM Study) also stresses observation and concepts. This study grew out of the recommendations of an ad hoc committee set up by the American Chemical Society to consider needed changes in high-school chemistry. The philosophy of the CHEM Study group is that the important concepts and generalizations of chemistry should be developed inductively and that the evidence that the students use should be gathered by them in the laboratory wherever possible. Where this is not possible it is presented by teacher demonstrations or films. There is emphasis on the unifying concepts in chemistry rather than on great volumes and varieties of technical applications.

The Biological Sciences Curriculum Study (BSCS) of the American Institute of Biological Sciences faced some problems that were more touchy than those faced by the physicists and chemists. Traditionally, the topics of evolution and of human reproduction had been taboo in high-school biology courses. The BSCS group felt that modern biology could not be taught without introducing these topics, and they, therefore, included them in the three courses which they developed. Though there has been some resistance to these controversial topics in a few places, there is now general acceptance of the new courses by schools throughout the country. Indeed, it is estimated that this year 80 percent of all students taking biology in the tenth grade are in one of the BSCS courses. Though the three BSCS programs take different approaches to the teaching of biology, the goals are the same – to present biology as a dynamic and growing field of knowledge, not as an encyclopedic catalog of facts. A major accomplishment of the BSCS group has been to change the high-school biology laboratory from a place where dead, preserved specimens are dissected to a place where living organisms are examined and experimented with.

In every case the curriculum groups saw to it that the materials were thoroughly tested in classrooms before being released for general use. Each of the groups prepared batteries of tests to determine whether high-school students could, indeed, learn science in these new ways. The results have been gratifying.

One of the important accomplishments of all of the curriculum groups has been to bring together scientists and teachers to work out the new programs. These are not courses handed to the schools by college professors, but are truly programs that have been developed cooperatively.

Physics has never been a popular subject for high-school students, and the introduction of PSSC physics has not stimulated more students to study this discipline. There has been concern among scientists and educators that college bound students, in general, should have some experience in this field. One group having this concern is developing a new high-school program which, it is hoped, will encourage non-science majors to take physics. The program, called Harvard Project Physics, adopts a cultural approach. It develops important concepts such as force, energy, and motion, but draws heavily on the humanistic background of the science. The program is still in an experimental stage, and it remains to be seen whether it will achieve its goal of increasing physics enrollment in high schools.

The immediate short-range goals of the completed high-school curriculum projects seem to have been reached in a satisfactory manner. Through tryout and evaluation it is clear that high-school students can learn science through the medium of the new programs. There are some areas, however, where further research remains to be done. Each of the curriculum groups either explicitly or implicitly suggest that their approaches should develop in students an understanding of and an ability to engage in the scientific enterprise, and an interest in and enthusiasm for science which should carry on beyond the high-school years. Little has been done to determine whether the programs do indeed have this effect. A long-range study to evaluate this aspect of the new curriculum programs would be desirable.

Another area which should be investigated is the effect of one program on another in a school. For example, one might ask whether a student in PSSC physics in the twelfth grade does better if he has previously had BSCS biology and CBA or CHEM Study chemistry than if he has not. The AAAS Commission on Science Education has plans for carrying out such a study but the plans have not been implemented.

A third and important question is one of sequence of the courses in high school. The curriculum groups designed their programs to fit into existing slots: biology, tenth grade; chemistry, eleventh grade; and physics, twelfth grade. This sequence of courses has become established over several decades but there is no logical reason for teaching the courses in this order. Indeed, when one looks at the interrelationships of the disciplines, logic suggests that the order should be inverted. Modern chemistry is deeply rooted in physics, and modern biology requires of the student more than a speaking acquaintance with chemistry. The groups who prepared the chemistry and biology curricula recognized the problem. The chemists were forced to include some of the ideas of physics - for example, the concept of energy – in their programs. The biologists in turn, had to include some simple chemistry in the hope that students would better understand some of the molecular biology which is included in the courses. The result in each case is not completely satisfactory, and it seems clear that some research needs to be done on this problem. I shall deal with this later. But now I want to turn to the pre-high school years and discuss the development of science curricula there.

PRE-HIGH SCHOOL

I said earlier that the new high-school science programs have set the pattern for science curriculum research in the pre-high school years. The pattern was set in several ways, including cooperation of scientists and teachers in producing the new programs, and emphasis on student activity and

and investigation rather than on didactic presentations. As the new high-school curricula came into being, the writers and the teachers who used them soon realized that many of the skills and attitudes that students need for effective participation in the new curricula should be learned in earlier grades. Let me take just one example. The scientists working with high-school teachers developing CBA chemistry soon learned from the teachers that they could not assume that students entering high-school chemistry had developed skills in constructing and interpreting graphs and so a section of the CBA program is devoted to developing these abilities. There seems to be little doubt that these are skills that could be developed in earlier grades.

Many of the scientists who worked on the high-school curricula are now devoting their energies, in collaboration with school teachers, to designing and testing science curricula for the early grades. There are so many of these programs for elementary and junior high school that we cannot examine them all in detail. But we will review most of the major ones starting with the programs for the early grades and telling you in some detail about the elementary science program on which I am working.

Each of the groups working at the elementary school level have the same general philosophy and plan of attack. The purpose of introducing new science programs into the early grades is not so much to recruit more scientists as to prepare individuals to be scientifically literate citizens. Literate not in the sense of knowing many scientific facts, but of understanding both the power and the limitations of science and its fundamental concepts.

The approaches that the various curriculum development groups are taking to achieve this goal vary widely. The Elementary Science Study of Educational Services, Incorporated takes the position that the child should be introduced to science by presenting him with a variety of materials to investigate. They have developed and tested in classrooms, units on butterflies, gases and airs, kitchen physics, microgardening, meal worms, and many other topics. Let us look at the meal worm unit as an example.

The children are given shoe boxes containing a few meal worms and are left to investigate them on their own. Gradually, as they observe them, they begin to ask questions. How do the meal worms find the edges of the box? How do they find their food? How do they back up? . . . and so on. As they design experiments to answer their questions, it is hoped that they will learn to evaluate evidence based on observation, classification, measurement, and control of variables; that they will learn to ask questions that they can answer by simple tests; and that, at the same time, they will develop an enthusiasm for scientific investigation.

The School Science Curriculum Project of the University of Illinois has a philosophy similar to that of ESS.

The Science Curriculum Improvement Study at the University of California takes a different approach. Its main thrust is to develop an understanding of science concepts starting in kindergarten with the simple concept of "What is an object?" and moving from there to "What is a system?", "What is meant by interaction?", and so on. Their program is more structured than ESS but still leaves room for investigation by children.

A science program for the early grades is the Minnesota School Mathematics and Science group, abbreviated "Minnemast." This group began by developing a mathematics program and then devising science units to go along with it. Their science program aims at developing both skills and concepts. So far they have seven units for kindergarten and first grade entitled "Watching and Wondering," "Describing and Classifying," "Our Senses," "Shape and Symmetry," "Objects and Their Properties," "Changing and Unchanging Properties," and "Introduction of Measurement."

The project with which I am working is developing a program, called Science – A Process Approach, for K through grade 6. This curriculum development is one of the activities of the AAAS Commission on Science Education. Shortly after it was established in the spring of 1962, the AAAS Commission on Science Education held two conferences of scientists, science educators, and teachers. From these conferences came the suggestion that a fruitful approach to teaching science in the early grades would be to devise exercises which would develop children's skills in using the processes of science. The Commission elected to carry out this task. This program, which is now being tried out in schools for the third year, is, like the other projects, a cooperative effort of scientists and school teachers.

A basic premise of this project is that a scientist's behaviors as he engages in his pursuit constitute a complex set of skills and intellectual activities but that these can be analyzed into simpler activities, and it is these simpler activities which we are calling the processes of science. Another premise is that the individual's ability to use these various processes can be developed in a step-wise fashion beginning in the earliest years of school. The processes for the early grades have been identified as observing, classifying, measuring, communicating, inferring, predicting, recognizing and using space/time relationships, and recognizing and using number relationships. For each of the processes there is a series of exercises of increasing complexity and difficulty which require the child to engage in activities in the classroom rather than to simply read or be told about science.

Let me give you a better idea of what the program is all about by describing in some detail one of the exercises written for the first grade. The exercise is the fourth one in the Communicating process and is entitled, "Introduction to Graphing." It starts with a statement of objectives – "Following this exercise the child should be able to":

1. construct a bar graph.
2. identify and name the number of items represented by the bars of such a graph.
3. state the number of items represented by the bars and distinguish between such expressions

as more than, fewer than, the same number as, most, and fewest.

There is a section called Rationale which tells the teacher why graphing is introduced at this level. It gives her some suggestions about contructing bar graphs and also includes suggestions of how to handle children's questions about why a graph is useful in communicating information.

The activities are introduced by putting into a box a number of colored blocks so that there will be one for each child and so that there will be at least three different colors. The teacher then asks each child to come and pick a block out of the box and then put it into a rough pile on the table so that the blocks of each color will be piled together. The teacher then asks the children if it is easy to find how many blocks there are of each color. The children, of course, find it necessary to come up to the table and count the blocks in the piles. The teacher then asks if they can think of a way of arranging the blocks so that the number can be found more easily and suggests, if the children do not think of it, that if the blocks are piled in vertical columns, they can be counted easily even from a distance. This leads ultimately to the question of ways of remembering how many blocks there are of each color and into the next activity in which bar graphs are first introduced. In that activity each child again selects a block and also a felt square and as he puts the block in one of the piles on the table he puts the felt square in one of the columns on a felt board. At the end of this activity there is a bar graph consisting of felt squares on the felt board, one column for each color of block. In the activities that follow the children make bar graphs by coloring squares on large sheets of paper and finally are introduced to the idea of a grid and of labeling the vertical axis of the grid with numerals and the horizontal axis with a key to identify the objects.

At the end of this exercise is an Appraisal which the teacher uses to determine whether the children have developed the skills set forth in the objectives. For the Appraisal the children are given objects such as toy vehicles of various sorts and asked to make graphs to indicate how many of each type there are and then to answer questions the teacher poses about the graph – questions such as: Are there fewer cars than motorcycles? Are there more trucks than motorcycles?

You will have noted that one of the processes that I mentioned is recognizing and using number relations. While there were some of the writers who felt that it was not necessary to include number exercises in a science program, the majority agreed that since number skills are required in many of the other process exercises, we should not simply assume that the child learned them in his regular mathematics program. The exercises in the numbers process are not intended to be a complete mathematics program but are intended to supplement the program already in use in the classroom.

At the fourth grade we begin to shift gears and introduce a new set of processes which are more clearly recognizable as activities that a scientist engages in. These processes are formulating hypotheses, controlling variables, experimenting, interpreting data, devising models, and making operational definitions. We sometimes speak of these as the integrated processes because they bring together the skills from the earlier part of the program.

One of the advantages of a program in which the objectives are stated in behavioral terms is that evaluation can be done by assigning the children tasks which will indicate whether or not they have attained the expected skills. For example, following the kindergarten exercise in which the children learn to identify and to name various plane shapes, such as circles, squares, triangles, and rectangles, it is a simple matter for the teacher to present a child with an array of plane shapes and ask him to point to a triangle, or the teacher might point to a circle and ask the child to name it. The results of each task can be recorded with a binary scale. Yes, the child could perform the task, or no, he could not. For each exercise we have prepared a series of tasks which we call the competency measure. Each teacher in our tryout program uses the competency measures on a sample of children from her class after she has taught each of the exercises. The results of this testing in the various tryout centers are collected and tabulated and used by the writers as they revise the exercises during each summer writing session. The results give the writers a good idea about the success or failure of any given exercise.

There are many aspects of the teaching in the classroom that we cannot learn about through the competency measure and so we ask each teacher to give her subjective impression of the exercise on a feedback form. This includes information about the time required to carry out the instruction, suggestions of activities that might be omitted or added, the reaction of the class, including interest and enthusiasm, carry over of learning to other activities or areas, and attention span. This feedback is also helpful to the writing group as the program is revised.

Although the emphasis in Science – A Process Approach is on developing skills in using science processes it is important to ask the question, "What about science content?" It is obviously impossible to teach process skills without using content of some sort, and the content that we use is drawn from various fields of science. To insure this we include in our writing team scientists from all major areas. In the early grades, of course, there are many exercises in which the relationship of the content to science appears to be somewhat distant. Learning to recognize and name two-dimensional and three-dimensional shapes is an example. Learning to make graphs using colored blocks is another. However, as the program progresses into higher grades the content is more and more clearly science. For example, among the twenty-five exercises in Part Five, which is used in fourth grade, are the following:

Inferring 8 – Inferences of Patterns in Electric Circuits
Defining Operationally 1 – Circuits, Conductors, and Nonconductors
Controlling Variables 1 – Growth of Mold on Bread

Interpreting Data 1 - Guinea Pig Learning in a Maze
Controlling Variables 7 - Orientation of Plants
Experimenting 1 - Separating Mixtures

Of the eighty-one exercises in Parts Five, Six, and Seven, the subject matter is approximately as follows: physics 33 percent; biology 21 percent; mathematics 17 percent; chemistry 9 percent; psychology and social science 9 percent; earth science 7 percent; astronomy 4 percent.

JUNIOR HIGH SCHOOL

We believe, then, that by the end of the sixth grade, children going through this program will not only have developed sufficient skills to carry out simple investigations but also will have a store of scientific knowledge which will make it possible for them to embark on a fairly sophisticated science program in the junior-high-school years. What the pattern for junior-high-school science will turn out to be is not yet clear. At the moment there are four groups who have developed curriculum materials for junior high school and tested them on a fairly wide scale. Then there are several other groups who are at earlier stages of curriculum development. The four programs that are farthest along are centered mainly in the physical and earth sciences and astronomy.

The astronomy program, known as the University of Illinois Elementary School Science Project, actually is designed for the later elementary grades and the early junior-high-school grades. This program is frankly subject matter oriented and consists of a series of childrens' books and teachers' guides. The program starts with the topic "Charting the Universe," which includes measuring distances in the solar system, measuring the size and shape of the earth, and using light as a tool for measuring great distances. The topics which follow it are "The Universe in Motion," "Gravitation," and "The Message of Starlight." Two additional topics which are still in preparation are "The Life Story of a Star" and "Galaxies and the Universe." These materials are not intended to serve as a complete science program for grades five through eight, but rather as topics to be included in a broader science program. The materials are being tested by several hundred teachers in various parts of the country but the results of the evaluation have not been reported.

The second program which includes some astronomy but many other topics as well, is the Earth Science Curriculum Project which is designed for ninth grade. This course was first written by a team of scientists and teachers in the summer of 1964, revised in the summer of 1965, and is already being prepared for commercial publication. The course, called Investigating the Earth, is an experience-centered course in which inquiry in the study of natural phenomena is stressed. It consists of four units – the first one, "The Dynamic Earth," considers the materials of the earth, earth changes, earth motions, force fields, and energy flow in earth processes. The second unit, "Earth Cycles," considers the hydrologic cycle, climate, and the earth's crust. The third unit, "The Earth's Past," discusses the measurement of geologic time and paleobiology. The course ends with a fourth unit, "Earth and the Universe," which begins with a discussion of the moon, moves on to the solar system and its origin, and finally to stars, galaxies, and cosmology. The program has been evaluated in eighty classrooms, and student progress has been determined with a test on science knowledge which is used both as a pre- and posttest.

A third junior-high-school program – Time, Space, and Matter – is being developed by the Secondary School Science Project at Princeton University. This program begins and ends in geology but includes a good deal of fundamental physical science as well. The students investigate topics such as solution, crystallization, and water erosion using well-designed equipment which is furnished to them in the form of individual kits.

A good deal of attention is being given to evaluating this program to provide feedback for those who are revising the course. Each teacher keeps a daily record of course progress and reports at regular intervals to the project office. The teachers are particularly requested to indicate where students show an extension of knowledge or interest to out-of-class situations and where they show signs of understanding or of confusion as they use the course materials. In addition, the group is developing a test to demonstrate changes in the behavior of children after they have had experience with the new curriculum. Other methods of evaluation, including the instructional interview technique, are being considered.

A fourth program, called Introductory Physical Science, is being developed by the same group that produced the high-school PSSC physics course. This course also is intended for ninth grade and is designed to develop in the students basic attitudes and skills that they will need when they later encounter the new high-school programs in physics, chemistry, and biology. The first two-thirds of the course considers matter and its properties and leads to the need to assume that matter is particulate in nature. The last third of the course considers the size of molecules, molecular motion, and energy. This course is laboratory centered, and the group has devised simple equipment which permits students to carry on significant experiments in ordinary classrooms.

It is clear as one looks at the new developments in science curricula for the junior high school that we are a long way from a smooth flowing science curriculum for these grades. There has been little or no coordination among the present curriculum development projects. As a first step toward coordination the AAAS Commission on Science called a conference of representatives of the junior-high-school curriculum projects in science and mathematics. The conference was held in July 1965 at Michigan State University. As the representatives of the projects reported on their work it became clear that there was a major gap in the junior-high-school curriculum developments. No group was working in the field of biology. The Biological Sciences Curriculum Study has, for some time, been planning to develop a biology program for junior high school, but they have made no progress since funds have not been made available to them.

At the conference, Bentley Glass and Arnold Grobman of BSCS suggested two possible types of junior-high-school science programs. One is a topical plan and the other a stratified plan. In the topical plan they suggested that the science program for grade 7 include about half a dozen topics, such as conservation of matter and energy, natural cycles, biological regulation, and homeostasis. Another half dozen topics would be developed for eighth grade. These would include change through time - cosmological, geological, and biological. Another group of topics would be developed for ninth grade and would include the broad aspects of meteorology, environment, and habitat.

The stratified plan that they suggested was as follows:

Grade 7 - physical science - a year course in descriptive aspects, largely earth sciences.
Grade 8 - life science - a year course in diversity and unity in the organic world, fundamental structure and function.
Grade 9 - physical science - a year course in basic concepts of matter and energy.

Another possible approach to an integrated junior-high-school science program was described at the conference by Dr. Robert Jastrow, chairman of the Junior High School Science Subcommittee of the AAAS Commission on Science Education. This program would follow naturally from the elementary program, Science - A Process Approach, but could also build on the other elementary science programs that I have already described. The course for seventh grade would serve as a transition between an elementary science program where the student has developed skills in using science processes and where he has learned some basic science content; and the eighth and ninth grades in which the course would develop a coherent picture of man and his physical and biological environment. The seventh-grade course, which would be rooted in laboratory experiences and field trips, would consist of selected topics from the physical and biological sciences and from astronomy and the earth sciences.

The eighth grade would begin with consideration of the basic materials and forces of the universe. The neutron, proton, and electron as fundamental units out of which all matter is constructed, and the basic forces of nature - gravity, electromagnetism, and the nuclear forces. The interplay of particles and forces would then be used in the discussion of the birth of stars and galaxies and the synthesis of chemical elements. The next step would be to consider current ideas about the formation of the solar system, nature of planets, and the probability that planets exist around other stars. Then geophysics and geology would be introduced in a discussion of the history and present structure of the earth and its atmosphere, and finally the course would take up the elements of molecular biology and combine this recently acquired biological knowledge with the geological discussion of the earth's history in broad attack on the problems of the origin of life and the biochemical foundations of evolution.

The ninth-grade course would treat biology from the point of view of evolution. The first topic to be considered would be the development of self replicating cells from primitive catalysts and complex organic chemical systems. Then multicellular organisms would be discussed and the similarities between plant and animal forms as far as capture, storage, and step-wise release of energy is concerned. Following this, differentiation and embryogenesis would be considered and then the molecular, organismic, and ecological factors involved in natural selection would be treated. The ninth-grade program would end with a discussion of the development of the nervous system and with how the nervous system accounts for consciousness, perception, learning, and memory.

A course of this sort, in which central problems of modern science serve as the core, should serve as a general education course for students who do not elect to take much science in high school. At the same time, it should provide a solid background of the skills and understanding which would serve as an effective base for the high-school biology, chemistry, and physics courses. No work beyond planning has been done on this program yet. It is hoped that work can begin within the next year.

SCIENCE COURSE SEQUENCE

Now I want to come back to the question about the sequence of science courses in high school that I have already mentioned briefly. One of the arguments for putting physics off until the twelfth grade is that the students do not develop the necessary mathematical skills until they have reached that level. I have not said anything about the new mathematics curricula except the relationship of mathematics to the Science - A Process Approach elementary curriculum. There are, of course, several new approaches to teaching mathematics in the schools. In these new programs the students develop mathematical skills at much earlier grade levels than they have done in the older traditional programs. It is possible, I think, that by the end of the ninth grade, students will have sufficient skill in algebra to do the mathematics in a modern physics course such as PSSC. Mathematics skill would, therefore, no longer be a block to moving physics to an earlier grade.

If a junior-high-school science curriculum gives the student some fundamental knowledge of physics, chemistry, and biology and related sciences, I see no reason why physics cannot be made a tenth-grade course, followed by chemistry in grade 11, and a molecular biology course in grade 12. Already, in a number of school systems around the country this sequence is being tried.

There is another possible pattern for grades 10 and 11. Since there are a number of topics that are common to physics and chemistry, it is possible to interweave these two disciplines into a two-year sequence of physical science. An experiment is now under way in Portland, Oregon, to do just this. A group there set out three years ago to devise a "road map" through PSSC physics and CBA chemistry and a similar one through PSSC physics and CHEM Study chemistry so that teachers could move easily from one book to the other. The program starts with the topic "Interpreting the Universe," and considers fundamental concepts and measurements and the nature of chemical change. The next section is mechanics which is all drawn from the PSSC course. The rest of the year is devoted to the electrical nature of matter, atomic models, and molecules and energy. In all of these chemistry and physics are interwoven. In the second year of the program there is less blending of the two subjects. The course begins with optics and waves, followed by electromagnetics and a discussion of quantum systems and then moves into consideration of bonds and chemical systems; order, disorder, and change; and molecules and energy.

This program has been evaluated using selected test items from CBA, CHEM Study, and PSSC tests. The Portland group reports that the combined program is successful and that even though students are permitted to take only the first year if they like, a large majority of them stay with it for the full two years. At present, this program is designed for grades 11 and 12, but it could be used at grades 10 and 11.

The "road map" technique is not completely satisfactory, and there is need for a fresh start to develop an integrated modern physical science curriculum.

REVIEW OF PROBLEMS

We have examined some of the recent and current science curriculum research and have suggested directions in which curriculum research might go in the future. There are some corollary problems that I have omitted. We will review these briefly.

As elementary science education moves from fairly passive learning through books and exposition to student manipulation and investigation, there will be an increasing need to provide supplies and equipment for classrooms even as early as kindergarten. The amount of money provided for science supplies in most school systems is small. It will have to be increased manyfold to accommodate the new programs. It is impossible to do more than estimate roughly what the cost of introducing the new elementary programs would be. It appears that on the order of $200 per classroom will be needed to introduce the new programs. To assist school systems in planning for the introduction of the new elementary science programs, the AAAS Commission on Science Education is just embarking on a study of the costs of equipment and supplies. It is our expectation that this study will be completed shortly.

Classroom design will also be influenced by the new science curricula. As a minimum for the early grades, classrooms should have multiple electrical outlets, running water, and level working areas. I have been surprised at the many classrooms I have visited that have only a single electrical outlet.

Although the new junior-high-school science curricula that are now under development are designed to be used in classrooms without special laboratory facilities, I think that in the future junior-high-school classrooms will need to be provided with laboratory space similar to that found in senior high schools.

A crucial problem raised by the new curricula is teacher training. As you know, the senior-high-school science teachers have been able to attend summer institutes to learn to use the new biology, chemistry, and physics courses. A few institutes are now being held for junior-high-school science teachers. For the elementary schools the problem is different both quantitatively and qualitatively. The science programs for the earliest grades do not require that the teacher know a great deal of science content, but they do require some teacher orientation. Up to now during the experimental stage the various curriculum groups have carried on this orientation and training on an ad hoc basis. In the future, as the programs are more widely used, school systems must establish inservice training programs. It would seem that high-school science teachers might be used to staff such programs.

And then there is the problem of preservice training of teachers. Only one of the curriculum groups, "Minnemast," is preparing college material to train prospective teachers to use the "Minnemast" materials. There is a group at Rensselaer Polytechnic Institute that is designing a new course in physical science for non-science majors with the thought that this course will better prepare new teachers to use the new school science curricula.

I said earlier that one of the important things that happened as the curriculum groups organized and developed their programs was that scientists and teachers collaborated, each learning from and educating the other. This cooperation has produced needed and significant changes in science curricula. Changes which could not have been effected by either scientists alone or teachers alone. This cooperation must, and I am sure, will, continue. Science curriculum research is a continuing activity, and the best minds in science and the best minds in education must be devoted to it.

Curriculum Research and Development in English

Robert C. Pooley
The University of Wisconsin
Madison, Wisconsin

THIS PAPER is planned to present selected items of recent research in the teaching of English from a number of aspects. Together they make up the body of information and expert opinion which will largely direct curriculum work in the decades ahead. Although the goal of most English teaching is to accomplish an integration of the component parts of the subject, they are here presented in divisions to avoid confusion and overlap. One area of research and study often associated with the English language arts is reading as a skill. It is omitted from the survey because in the field of education it is considered separate from the language arts. It has its own courses, its own facilities, and a very extensive body of research, only a small portion of which is related to the language arts. The word "research" has been interpreted broadly enough to include mention of seminal books which, if not in themselves research, have induced and will induce research in the areas of their treatments.

THE PROFESSION OF ENGLISH TEACHING

In 1961 a committee of the National Council of Teachers of English published a document entitled The National Interest and the Teaching of English with the sub-title A Report on the Status of the Profession. In three major parts this document studied the present situation in the teaching of English, related the situation to the national scene, and made analyses and studies leading to specific recommendations for the improvement of the teaching of English by meeting recognized needs.

The scope of the responsibility of English is dramatized by a bar chart revealing that of all high school students in school in 1960, 92.9 percent were enrolled in English courses, 68 percent in social sciences, 55 percent in mathematics, and 25 percent in foreign languages. A survey revealed that in the years 1958-1959 the demand for trained high-school teachers outran the supply by 27 percent. Expanding school populations will increase this ratio. Moreover, the expansion of knowledge and the requirements of a complex society will increasingly demand more technical training and knowledge on the part of English teachers than was true in the past. The single greatest weakness in the current situation is the lack of articulation in the teaching of English from the elementary school through college.

A summary of the highlights of the status report follows:

The Need for Teachers of English:
  a. The shortage of teachers shows no sign of decreasing.
  b. One-fourth of all elementary teachers are not college graduates.
  c. Only 40 to 60 percent of high-school English teachers have completed a college major in English.
  d. There is an acute shortage of college instructors in linguistics and composition.

State Certification Laws:
  a. States differ widely in requirements.
  b. Ten states certify elementary teachers without a degree.
  c. Nineteen states have no stated requirement in English for elementary teachers; twenty-

one lack definite requirements in reading, children's literature, and English methods.

- d. Minimum preparation in English for English teachers is sixteen to eighteen credit hours.

Preparation of elementary teachers:

- a. More than 94 percent of the colleges fail to require specific study of the English language.
- b. Less than half require a course in the teaching of the language arts.

Preparation of secondary-school teachers of English:

- a. Only one-fourth of the colleges require a course in the history of the English language.
- b. Only 17 percent require a course in Modern English Grammar.
- c. Only 41 percent require a course in advanced composition.
- d. Only slightly more than half the colleges require a course in the methods of teaching English.

Better Teaching Conditions:

Teachers with a student load of 150 (five sections of thirty students) must work more than 50 hours a week just meeting classes and marking papers. (In 529 out of 682 schools, English teachers had five classes.)

Libraries:

- a. 10,000,000 elementary, and about 150,000 high-school students attend schools lacking a central library.
- b. The average expenditure for books per student is about half the price of one average book.

Remedial English in Colleges:

- a. An estimated 150,000 students failed college English tests in 1960.
- b. Two-thirds of American colleges have to offer remedial work in English.
- c. The cost of remedial instruction is estimated to be over ten million dollars.

A continuation of the research into the status of the teaching of English was carried on by a committee of the National Council of Teachers of English and published in 1964 under the title The National Interest and the Continuing Education of Teachers of English. The principal purpose of this investigation was to discover and reveal the need for inservice education of teachers concerned with English in elementary and secondary schools. Questionnaires were answered by 7,417 secondary-school English teachers and by 3,030 elementary-school teachers.

Some highlights of the analysis are:

1. Over 40 percent of the nation's elementary teachers began full-time teaching without a degree. Only 31 percent of these achieved the degree after beginning teaching.
2. More than one in ten high-school teachers (11.8 percent) began full-time teaching without a degree. Practically all of these earned a degree later.
3. Although 24 percent of instructional time in kindergarten through grade 12 is given to English and language arts, only 8 percent of the elementary teachers' college preparation is given to this subject.
4. About half (49.5 percent) of secondary teachers teaching English lack a major in this subject.
5. By their own admission only 51.9 percent of secondary-school English teachers feel competent in literature; 36.6 percent feel competent in composition, and 53.5 percent feel competent to teach the English language.
6. In voluntary continuing education, elementary teachers take one summer session for each eight years of experience. But the courses taken are largely in education, with little or no further education in the language arts.
7. Thirty percent of secondary teachers of English have not taken a course in English in more than ten years.
8. Interest in further education for elementary teachers was expressed for six areas: children's literature, language, speech, writing, general literature, and the largest number, 79.1 percent in reading.
9. Secondary teachers expressed interest in further experience in practical methods of teaching English: intermediate or advanced composition, literature for adolescents, teaching of reading, literary criticism, and structural and/or generative grammar.

The broad conclusion is overwhelming evidence of inadequate training with a clear need for means to make inservice further education available to nearly all teachers of the English language arts.

Two books of importance to the whole profession of the teaching on English appeared in 1963 and 1965 respectively. The Education of Teachers of English (Grommen, 1963) is Volume V of a curriculum series of the National Council of Teachers of English. Written by a committee representing elementary-school, high-school, and college English teaching, it analyzes facts, implications, and inferences in such matters as the liberal education of teachers, the professional education of English teachers, and the specific preparation in English itself. It confirms the conclusions reached in other studies that elementary-school teachers are very poorly prepared to teach English; that the preparation of high-school teachers is subject to wide variations and some serious omissions; and that college teachers, though perhaps well grounded in content, are often innocent of any preparation to teach. Much space is devoted to recommendations and procedures for the continuing education of teachers by means of summer sessions, institutes, evening classes, and uses of mass media.

The second broad area book is Freedom and Discipline in English, a report of the Commission on English of the College Entrance Examination Board (1965). The Commission on English was appointed in September 1959. Consisting of sixteen members from universities, colleges, and private schools, it undertook to define the field of English, and to set standards for instruction in English and in the preparation of teachers for English. In 1962 it sponsored and supported twenty institutes attended by 868 teachers in all parts of the United States. It commissioned and issued a series of kinescopes featuring experienced college teachers of English expounding their specialties. These kinescopes are lent free of charge except for mailing costs. The volume Freedom and Discipline in English is the final report of the Commission.

Chief contributions to the curriculum are:

1. Recommendations to improve the quality of instruction.
2. An analysis of language teaching, including grammar and usage.
3. An analysis of literature teaching, emphasizing the critical process in teaching and studying literature.
4. An analysis of composition, including valuable suggestions for motivating writing, and for the correction, criticism, and measurement of writing skill.

Without essential conflict in fundamental matters, it may be stated that, in general, the publications of the National Council of Teachers of English tend to be experimental and progressive in outlook, and that the point of view of the Commission on English is conservative. To the curriculum director there is great value in the balance to be derived from these differing viewpoints.

RESEARCH IN THE TEACHING OF LITERATURE

Although most of the studies in literature deal with ideas and values rather than with controlled evaluation, some recent studies reveal effective techniques for measuring the success of students in the learning of kinds of literary materials. Squire (1963) studied the responses of adolescent readers to four short stories, establishing their responses as forming seven categories: I. Literary Judgments: Direct or implied judgments on the story as an artistic work; II. Interpretational Responses: Reactions in which the reader generalizes and attempts to discover meaning, motivation, and the nature of the characters; III. Narrational Responses: Reporting facts or details from the story without interpretation; IV. Associational Responses: Responses in which the reader associates ideas, events, places, or people with his own experience; V. Self-involvement: Responses in which the reader associates himself with the behavior and/or emotions of the characters; VI. Prescriptive Judgments: Responses in which the reader prescribes a course of action based on some absolute standard; VII. Miscellaneous Responses: all others.

Wilson (1966) reports the responses of college freshmen to the reading of three novels: The Catcher in the Rye (Salinger), The Grapes of Wrath (Steinbeck), and A Farewell to Arms (Hemingway). Fifty-four freshman English students read a novel, took notes while reading, and wrote immediate reactions in class. After three discussion periods they wrote follow-up reactions. This procedure was followed for each of the three novels.

The written responses were coded by means of Squire's (1963) categories. Statistical analysis showed that 65. 6 percent of responses to the three novels were interpretational; 12. 2 percent literary judgment; self involvement, 8. 9 percent; narrational, 8. 7 percent; associational 2. 1 percent; prescriptive 1. 3 percent; miscellaneous 1. 3 percent. The lead of interpretation over literary judgment is a sign of maturity of perception; it suggests that teachers should avoid supplying "correct" interpretations before the book is read by the student. The study reveals that although students may begin their involvement with literature in a groping and emotional fashion, their later attempts at interpretation become sharply formulated and logical. Teachers should therefore refrain from demanding that first reactions be fully described or defended. But as the study of the novel moves into the grappling with difficulties, the instructor should demand expert analysis. Most students are capable of growth of interpretation by these means.

The Research Foundation of the National Council of Teachers of English is sponsoring, directing, and supporting research in children's growth in sensitivity to, and appreciation of literature. Starting with a national conference of elementary-school, high-school, and college specialists in literature assembled in Chicago in May 1963, the Foundation applied itself to research in the teaching of literature because so little is being done. It was soon discovered that there were no instruments available to measure any aspect of success in children's growth in literature. The Foundation then determined to try to create such tests, both as research in itself, and as the means to further studies. Following consultations with various authorities, the Foundation appointed Margaret Early of Syracuse University the Director of the test activity and appropriated funds in its support. In the summer of 1965, Early assembled at Champaign, Illinois, some twenty-five selected teachers and librarians to develop test items. From the materials produced by this group, and augmented by Early's advisors, four forms of a test have been developed by cooperative work with the Educational Testing Service of Princeton, New Jersey. A trial run was made of these forms in the spring of 1967 at elementary schools in grade 5 in locations covering the United States. From the statistical study of the returns it is expected that two final forms of the test will be developed.

The goal is to attempt to discover growth in the responses of children to literary experiences. The range is from grade 4 to grade 6, inclusive. Although it will be a pen and paper test, many of the items will be presented orally on tapes. If the test reveals measurable growth in responses to literature, the research will continue in two directions: (1) to measure the effectiveness of kinds of instruction on the literary growth of children; and (2) to create similar tests for junior-high schools and possibly senior-high schools.

RESEARCH IN THE TEACHING OF COMPOSITION

One of the most exhaustive researches of recent times was conducted by Richard Braddock, Richard Lloyd Jones, and Lowell Schoer of the University of Iowa and published in 1963 by the NCTE under the title Research in Written Composition. It undertook to assess all known research in composition that could be described as "scientific." From twenty bibliographies of research came over a thousand citations; committee analysis reduced these to 485 items. From consultation with seventeen authorities this number was reduced to about one hundred titles; later some fifty other titles were added. After vigorous analysis, only five studies survived the criteria set up by the investigators. These five only, in the eyes of the investigators, can be termed valid in the teaching of composition.

1. The Baxton study revealed that thorough criticism of composition by a competent reader, followed by revision by the student brought better gains than writing not criticized.

2. The Harris study showed in a 2-year space that the direct (or functional) teaching of grammar brought better results than the formal, analytical method.

3. The Kincaid study undertook to reveal whether a single paper written by a student on a given topic at a particular time provided a valid means to judge a student's writing ability. Joined to this question were several others, so that clear answers were almost impossible to secure because of weakness in the structure of the research.

4. The Smith study dealt with class size as a factor in the teaching of writing. It revealed that a large class of superior ninth-grade students can progress in composition as well as a small class.

5. The Becker study compared three methods of teaching, described as normal, bibliographic, and kinescope methods. The so-called normal method was found to have a slight superiority over the others.

The conclusion of the Braddock investigations is that of the more than a thousand items originally noted, only these five may be considered to yield valid answers from sound research.

In a review of the Braddock report in College English, Jean Hagstrom (1964) analyzes the research methods employed in these five researches, and concludes, "It is therefore extremely disheartening to say that 1) none of them strikes a layman as definitive or persuasive, and 2) there is very little promise that without rigorous antecedent thought, the 'scientific' method applied to composition will yield better results in the future than it has in the past." He urges antecedent thought on such basic problems as: How do you find a subject? How do you clothe it in appropriate style? How do you give it effective organization?

The Research Foundation of the National Council of Teachers of English planned a 1967 conference of specialists in a number of fields to attempt to identify the fundamentals of composition.

Current research in composition has taken an interesting and profitable direction in applying statistical procedures to rhetoric. Kellogg W. Hunt (1965a, 1965b) collected large samples of students' writing at grades 4, 8, and 12. These were subjected to structural analysis to attempt to describe the differences in sentence structure as students mature. As a tool of analysis Hunt developed what he called the "T-unit": a group of words minimal in length, each terminated grammatically between a capital letter and a period. This unit has the advantage of preserving all of the subordination achieved by a student, and all of his coordination between words and phrases and subordinate clauses.

His findings, very much condensed, are shown in Table 1. This table reveals that the most significant mark of maturity in writing is not sentence length in itself, but the increase of length of the T-units. When superior adults writing for Harpers and the Atlantic Monthly were analyzed by the same means, the length of their T-units increased by 40 percent over the twelfth graders', an increase equal to the gain of twelfth graders over fourth graders. Twelfth-grade students came close to the professionals in the number of clauses they use; the growth among professionals is in the increased number of words per clause. These results give a clear indication of the factors to be studied and taught in the improvement of sentence structure.

Francis Christensen (1963b) counted sentence openers to provide objective knowledge of how professionals begin sentences. High-school and college composition teachers have generally encouraged a wide variety of sentence openers, including prepositional phrases of all types, participial phrases, gerund phrases, absolute constructions, and sentence inversions. One writer says, "When a student opens a sentence with an infinitive phrase or a past participle, we immediately stamp him as more mature. . ."

Christensen analyzed twenty recent or contemporary writers – ten of narrative and ten of expository writing – to determine by what constructions they open sentences. The findings are in direct opposition to the traditional teaching: the professionals place something before the subject in only a fourth of their sentences (24.47 percent); that is, over 75 percent of their sentences begin with the subject. Of those opening by something other than the subject 23 percent of the total sentences open with adverbial modifiers, 1.17 percent use verbal groups, and the inverted structure appears in only one of every three hundred sentences (0.32 percent of the total sentences). The excessive use of verbal groupings as openers produces a style which Christensen calls "pretzel prose"; that is, twisted sentences. He concludes: "The good teacher should not only base his teaching on the practice of professional writers, but should himself practice before he preaches."

In another study Christensen (1963a) develops what he calls "A Generative Rhetoric of the Sentence." He presents a scheme for the purpose of "buckling down to the hard work of making a difference in the student's understanding and manipulation of language." From his own analysis and that of others he draws four principles underlying the rhetoric of the sentence:

1. Composition is essentially a process of addition.
2. The cumulative sentence requires direction of modification or direction of movement.
3. Additions are made in levels of generality or levels of abstraction.
4. Density of texture is achieved by adding frequently and much to nouns, verbs, and main clauses.

TABLE 1

MARK OF MATURITY IN WRITING

	Average length of clauses	Ratio of clauses per T-unit	Average length of T-units	Ratio of T-units per sentence	Average length of sentences
Grade 4	6. 6 words 77 percent	1. 30 77 percent	8. 6 words 60 percent	1. 6 137 percent	13. 5 words 80 percent
Grade 8	8. 1 words 94 percent	1. 42 85 percent	11. 5 words 80 percent	1. 37 117 percent	15. 9 words 94 percent
Grade 12	8. 6 words 100 percent	1. 68 100 percent	14. 4 100 percent	1. 17 100 percent	16. 9 words 100 percent

Using visual analysis Christensen shows how sentences are built (or generated) by the process of coordination (addition of parallel elements), by subordination, (addition of levels of modification) or by a combination of coordination and subordination. In essence Christensen recommends and demonstrates the superiority of the cumulative sentence over the complex sentence. His technique of analysis can enrich the study of literature as well as develop mature writing.

Two years later Christensen extended his theory of the sentence to the study of the paragraph. Using a parallel title "A Generative Rhetoric of the Paragraph" (Christensen, 1965), he shows that the principles of analysis he developed for sentences apply with equal cogency to the paragraph; that is, paragraphs are created by addition, involving coordination of elements, subordination of elements, or combinations of the two. As well as addition, there must be also direction of movement; there will be levels of abstraction, and varying density of texture. In applying these principles he formulated them under nine headings, a few of which are:

1. The paragraph may be defined as a sequence of structurally related sentences.
2. The top sentence of the sequence is the topic sentence.
4. Simple sequences are of two sorts: coordinate and subordinate.
6. Some paragraphs have no top, no topic, sentence.

The nine headings are developed by exposition and illustration. The conclusion to be drawn is that the paragraph may be studied and taught as the sentence is studied and taught by Christensen.

A. L. Becker (1965) offers a different analysis of the paragraph under the title "A Tagmemic Approach to Paragraph Analysis." A Tagmeme is defined as the shortest grammatical unit; Becker calls them spots or slots in a system where substitution is possible. They are composites of form and function. Limiting his analysis to expository paragraphs, Becker finds two fundamental patterns. The first has three functional slots which he labels T (topic), R (restrictions), and I (illustration). These slots correspond to three levels of generality in paragraphs (cf. Christensen) and reflect a natural way of talking or writing about something.

The second major pattern has two slots labeled P (problem) and S (solution). The P may be a question or an assertion; the S is the answer or response. Where the S is extended, it tends to follow the TRI structure. The majority of expository paragraphs fall into these two basic patterns.

Variations in the patterns are caused by deletion, reordering, addition, and combination. The formal markers of internal tagmemic structure are graphic (indentation), lexical (equivalence classes and lexical transitions), grammatical (changes in the grammatical role of equivalence classes signal new slots or new paragraphs), and phonological (shifts in pitch, register, tempo, and volume in oral reading).

This analysis makes possible the partitioning of paragraphs in a consistent and predictable way.

The accurate and impartial evaluation of composition is one of the major problems of English teaching. Paul Diederich (1966) made a study of the factors that account for differences in the grades of qualified readers. Three hundred papers on two subjects were read by sixty distinguished readers representing six different fields: English, social studies, natural science, writers and editors, lawyers, and business executives. By comparison of the ratings of three readers who stood highest in a particular factor with three who stood lowest in the same factor, it was possible to isolate influential factors. In all, 11,018 comments on 3,557 papers were tabulated.

The factors revealed stood in this order of frequency: ideas, mechanics, organization, wording, style or "flavor." For experimentation in high schools, these were divided into two groups: I. General Merit (ideas, organization, wording, and flavor); II. Mechanics (usage, sentence structure, punctuation, capitals, abbreviations, spelling, hand-

TABLE 2

DISTRIBUTION OF RATINGS

	Non-Academic (N = 476)		
Grade	10 (145)	11 (167)	12
High	5 percent	8 percent	9 percent
Middle	34 percent	53 percent	63 percent
Low	61 percent	39 percent	28 percent

writing, neatness). A rating card was developed for the use of evaluators. A close following of the techniques developed by Diederich, with a year of practice for a faculty, should yield a reliability of the cumulative total for each student close to .90.

In the same article Diederich (1966) develops an elaborate but practical procedure for the rating of large numbers of papers. By a process of sorting and grouping of anonymous papers of combined grade groups, such as grades 10, 11, and 12, the process results in precentage figures for selected groups like Non-Academic and Academic with significant differences for High, Middle, and Low in each group, and a recognizable growth figure from grade 10 to grade 12. This extract (see Table 2) from a senior-high population will illustrate the distribution of ratings.

In this lower ability group the progress in the highest level is slight but observable: 4 percent. The progress in the middle level is striking: from 34 percent to 63 percent; the reduction in the low group is equally striking: from 61 percent to 28 percent. The figures from the Academic group show great advances in the high and middle level. This procedure demonstrates to teachers, administrators, and parents a positive progression of improvement not demonstrable by customary paper-grading processes.

RESEARCH IN LANGUAGE LEARNING

Ruth Strickland (1962) undertook to analyze the structure of children's oral language in the first through the sixth grades, to compare it with the structure of language found in the texts designed to teach children to read, and to discover, at selected grade levels, the influence of any ostensible differences in the quality of children's reading skill. A major hypothesis was that a study of children's speech might offer suggestions for the construction of better textbooks in reading.

In the analysis of structure, twenty-five consecutive sentences recorded on tapes were used as the language sample for each child. A special scheme was devised for the analysis of the language at two levels. Level One was divided into "slots" or stationary elements, and "moveables," or elements which can appear in different positions. Level Two was concerned with the types of subordinations or "satellites" used in the fixed slots and the moveables. This structure analysis is an important contribution of the of the study. Tabulations from this analysis were programmed to computers to relate the frequencies to the variables of age, sex, intelligence, socio-economic status, and parent education. It was found that many similar patterns were used with great frequency by children of every grade. It was also noted that children of all levels used a wide range of language patterns, and could elaborate and expand sentences by the use of moveables and subordinate elements.

In the comparison of children's language with that of textbooks in reading, four sets of widely used textbooks were analyzed for language patterns. The patterns used by children and by texts at relevant grade levels were studied. No scheme for control over sentence structure to parallel that over vocabulary seemed to exist. This raises the question: Can knowledge of the structures used by children be used to devise textbooks controlling the difficulty of reading?

A study of the relationship between children's use of language and the skill they demonstrate in learning to read in grade 6 revealed that children ranking high in the variable elements made more use of the common oral structural patterns than did children ranking low; the same children also made greater use of moveables and elements of subordination than did children ranking low. Among the conclusions to be drawn are these:

a. Children's command of language patterns is much broader than is generally realized.
b. There is almost no relationship between children's oral patterns and the patterns of structure in reading texts.
c. Children can be helped to recognize the patterns of expression and by this means be aided to advance in reading and in the learning of grammar and usage.

Walter Loban, of the University of California, began in 1952 a longitudinal study of children's growth in language which continued for eleven years. His report The Language of Elementary School Children (Loban, 1963) covers the first seven years, the elementary school level. From the Oakland, California schools 338 subjects were chosen at kindergarten level and were observed and tested in the following years. Items especially studied were: their vocabulary, their use of oral and written language, their proficiency in reading and listening, teachers' judgments of their skill with language, and background information on their health and homes.

Loban developed what he calls "the communication unit" in analyzing children's speech. It may be defined as a group of words which cannot be further divided without the loss of their essential meaning. Another term he developed is "the maze," a series of words or initial parts of words which do not add up to meaningful communication. Language samples recorded on disks and tapes provided the source materials. Two sub-groups, one exceptionally high and one exceptionally low in language ability, were identified.

A large number of specific findings have been summarized in the published report as follows:

> During the first seven years of schooling, children speak more words in each succeeding year, produce more communication units, and increase the average number of words in those units.
>
> Children rated as skillful in language reduce both their incidence of mazes and the number of words per maze.
>
> Not pattern, but what is done to achieve flexibility within the pattern proves to be a measure of effectiveness and control of language at this level of language development.
>
> Children who are rated as most proficient in language are also those who manifest the most sensitivity to the conventions of language.
>
> Those high in general language ability are also high in reading ability. . . Writing ability is related to socioeconomic position . . . The highest correlation in the study is between vocabulary and intelligence . . . A low but positive relation exists between health and language proficiency . . . Competence in the spoken language appears to be basic for competence in reading and writing.

An important aspect of language learning in children is their ability to learn and use standard English. Loban in Problems in Oral English (1966) states that for most children a command of standard English is necessary, for "society exacts penalties of individuals who do not possess it." In continuing his longitudinal study of children's language, Loban here examines examples of nonstandard oral usage among four different groups of school children from kindergarten through grade 9. His chief findings have been summarized thus:

> In ten years of schooling, pupils from homes in which class dialect is used make almost no improvement in using the verb to be appropriately or in standardized verb forms.
>
> The practice, so common among weaker teachers, of drilling all pupils on the same skill is not supported by this research. Individual pupils, but not whole classes of pupils, will need help if they are to use the standard forms of irregular verbs.
>
> As the complexity of sentence structure and total volume of spoken language both increase, there is a more than proportional probability of difficulty with . . . problems of clarity and precision, not problems of habitual usage.
>
> Almost all the pupils whose parents speak informal standard English have little need of drill on usage.
>
> Oral drill is more effective than workbook drills.

THE EMERGING GRAMMARS AND DIALECT STUDY

In the last fifteen years a fundamental tenet of English teaching, namely, that there is one grammar of English to be learned and taught, has been blasted by the development of at least two other systems of English grammar and the reasonable expectancy of others to follow. It is no longer accurate to speak of the grammar of English except in reference to the complete operation of the English language. We now have several grammars, each a system to describe how the English language works, or more specifically, how English makes sentences. The traditional system was developed principally in the eighteenth century by scholars whose grammatical disciplines were derived from Latin and Greek. Traditional grammar, therefore, was developed to make English conform to structural principles of Latin and Greek. Aside from certain classical biases, and lack of then undeveloped linguistic principles, it is a system of analysis of English of great value in the past and is by no means obsolete at present, unless viewed as the grammar of English.

In 1952 C. C. Fries introduced structural grammar. He was chief among a number of scholars working along parallel lines. The contribution of structural grammar is accurate description: it is concerned with the detailed elements of a language, its sounds, word units, inflections, and syntax. In the latter, analysis is conducted on the basis of immediate constituents as operating in modification, predication, complementation, and subordination. Structural grammar has greatly enriched knowledge about the elements of English and the manner in which they are combined to express meaning.

Noam Chomsky (1957) of MIT brought out in published form a new approach to grammar under the title Syntactic Structures. Somewhat modified by revisions, this newest of grammatical theories has great promise for the future. Generally called generative-transformational grammar, it is a system which attempts to state the "rules" which generate all English sentences and only English sentences. It assumes that we have in us a mechanism for using these laws which we gain from intuition. We are therefore able to generate sentences that have never been uttered before. The analysis begins with kernel sentences, simple declarative sentences without development. From these kernels other sentences are derived by means of rules for transformations, rules which change or develop kernel sentences containing phrase structures which are grammatical parts of a grammatical sentence. The goal of this new grammar is to state the rules that will make the production of sentences predictable.

Applications of structural and transformational grammar for classroom purposes are developed by Owen Thomas (1965) and Paul Roberts (1964). The former develops the transformational system in easy stages for the English teacher to understand and utilize the processes of generative-transformational grammar in the instruction of students. The latter presents a set of programmed lessons for self-instruction in the new grammar. N. S. Blount (Blount and Johnson, 1966) of Madison, Wisconsin, has created a similar programmed series of lessons adapted to the eighth grade.

The following curriculum question arises: "What happens to students who are taught the structure of English by generative-transformational processes?" Bateman and Zidonis (1966) offer a tentative answer as the result of an experiment with ninth- and tenth-grade students. Using equivalent experimental and control groups, they taught the former by transformational techniques and the latter by traditional techniques. A valid sampling of free writing of each student was collected at the beginning and end of the instruction. The analysis of these writings by the Hunt (1965b) process revealed that the experimental group wrote better sentences of more sophisticated structure than did the control group. A parallel experiment is now under way at Manitowoc, Wisconsin, under the leadership of N. S. Blount of The University of Wisconsin, Madison.

A *Dictionary of American Regional English* is in progress, centered at The University of Wisconsin and directed by F. G. Cassidy of the Department of English (1966). Such a dictionary has been planned for over seventy-five years, and active collecting has been going on since 1889. Now under a 5-year contract with the U. S. Office of Education the work is progressing rapidly. Collections of data for six states, including Wisconsin are completed; six more are partially completed. Work of a similar nature is well under way in eleven more states. Many of the field workers live and work in "Word Wagons," specially equipped small buses containing the essentials of living quarters and apparatus for field collection.

Data are assembled, classified, and placed in a computer file at Madison. Collectors record speech; readers analyze regional literature and local newspapers. The effort is to secure complete coverage of the words and expressions used *differently* in various areas of the United States.

Curriculum applications of this study have yet to be worked out, but they will include 1) greater and more precise knowledge of American dialects; 2) a classification of usage levels and standards; and 3) new attitudes toward local and regional language variations.

WISCONSIN ENGLISH CURRICULUM ACTIVITY

By means of a Federal contract initiated in May, 1963, the State of Wisconsin has been served by the Wisconsin-English-Language-Arts Curriculum Project, jointly sponsored by The University of Wisconsin and the Department of Public Instruction. By means of two series of bulletins, conferences, conventions, and personal school visits, more than 5,000 elementary-school and high-school English teachers have been involved, many very actively. The project is, in fact, a statewide in-service training program in the English language arts. The aim is to encourage teachers to think, confer, discuss, and write about the curriculum in English. They have responded magnificantly. Over 3,000 bulletins go out each month; in the larger cities these are duplicated and distributed to additional numbers of teachers. By means of the reports of group leaders the project director is informed of the number of teachers who are meeting, the grade levels they represent, and their views on the issues that are raised.

In the summer of 1964 a selected group of teachers representing all school levels met to produce a curriculum in the teaching of literature – a cumulative growth pattern from the kindergarten through grade 12. After some months of revisions and editing, this was published in January of 1965 as *Teaching Literature in Wisconsin*, a paper-bound volume of 160 pages. Nearly 5,000 copies were distributed free in Wisconsin; it is estimated that the total distribution, including sales at cost, is about 17,000 copies.

By a similar process in the summer of 1965 a new group of selected teachers prepared the materials for the volume *Teaching Speaking and Writing in Wisconsin* (1966). Over 4,000 copies were distributed free; a large number have been sold at cost.

Again in the summer of 1966 two curriculum conferences were held, the first composed of ten selected college or university teachers who used their linguistic knowledge and experience in the teaching of English to plan the language curriculum. This was followed by a group of teachers representing elementary, junior high school, and senior high school levels which was to work out a continuum of language learning from kindergarten through grade 12. The result of this creative work and subsequent editing was issued in February of 1967 under the title *Teaching the English Language in Wisconsin.* Featured in this volume are:

1. A linguistic approach to grammar, utilizing simple applications of structural and transformational grammar to the language learning of children and youth.
2. A program that is sequential and non-repetitive.
3. A grammar that provides for more sophisticated sentence analysis at the time students need it.
4. An enrichment of the language program by the inclusion of word-study, semantics, and the history of the English language.

The three curriculum volumes listed above are issued in experimental editions. They were thoroughly revised in the light of experience in the classroom and reissued in the Spring of 1968 as one volume entitled *English-Language-Arts in Wisconsin.*

REFERENCES

Bateman, D., and Zidonis, F. K. *The effect of a study of transformational grammar on the writing of ninth and tenth graders.* Research Report No. 6. Champaign, Illinois: National Council of Teachers of English, 1966.

Becker, A. L. A tagmemic approach to paragraph analysis. *College Composition and Communication,* 1965, *16,* 237-242.

Blount, N. S., and Johnson, Shelby L. Selected concepts in structural and transformational grammar. Twenty-two unpublished lessons, Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, 1966.

Braddock, R., Lloyd-Jones, R., and Schoer, L. *Research in written composition.* Champaign, Illinois: National Council of Teachers of English, 1963.

Cassidy, F. G. Dictionary of American regional English. Unpublished progress report, 1966.

Chomsky, N. Syntactic structures. The Hague, Netherlands: Mouton, 1957.

Christensen, F. A generative rhetoric of the sentence. College Composition and Communication, 1963, 14, 155-161. (a)

Christensen, F. Notes toward a new rhetoric. College English, 1963, 25, 7-18. (b)

Christensen, F. A generative rhetoric of the paragraph. College Composition and Communication, 1965, 16, 144-156.

College Entrance Examination Board. Freedom and discipline in English. Princeton, New Jersey: CEEB, 1965.

Diederich, P. B. How to measure growth in writing ability. English Journal, 1966, 55, 435-449.

Fries, C. C. Structural grammar. New York: Harcourt, Brace and World, 1952.

Grommen, A. (Editor) The education of teachers of English. New York: Appleton-Century-Crofts, 1963.

Hagstrom, J. H. Review of National Council of Teachers of English, Research in written composition. College English, 1964, 26, 53-56.

Hunt, K. W. A synopsis of clause-to-sentence length factors. English Journal, 1965, 54, 300-309. (a)

Hunt, K. W. Grammatical structures written at three grade levels. Research Report No. 3. Champaign, Illinois: National Council of Teachers of English, 1965. (b)

Loban, W. The language of elementary school children. Research Report No. 1. Champaign, Illinois: National Council of Teachers of English, 1963.

Loban, W. Problems in oral English. Research Report No. 5. Champaign, Illinois: National Council of Teachers of English, 1966.

National Council of Teachers of English. The national interest and the teaching of English: A report on the status of the profession. Champaign, Illinois: NCTE, 1961.

National Council of Teachers of English. The national interest and the continuing education of teachers of English. Champaign, Illinois: NCTE, 1963.

Roberts, P. English syntax. New York: Harcourt, Brace, and World, 1964.

Squire, J. R. The responses of adolescents while reading four short stories. Research Report No. 2. Champaign, Illinois: National Council of Teachers of English, 1963.

Strickland, Ruth G. The language of elementary school children. Bulletin of the School of Education, Indiana University, 1962, 38, No. 4.

Thomas, O. Transformation grammar and the teacher of English. New York: Holt, Rinehart, and Winston, 1965.

Wilson, J. R. Responses of college freshmen to three novels. Research Report No. 7. Champaign, Illinois: National Council of Teachers of English, 1966.

Wisconsin English Language Arts Curriculum Project. Teaching literature in Wisconsin. Madison, Wisconsin: Department of Public Instruction, 1965.

Wisconsin English Language Arts Curriculum Project. Teaching speaking and writing in Wisconsin. Madison, Wisconsin: Department of Public Instruction, 1966.

Wisconsin English Language Arts Curriculum Project. Teaching the English Language in Wisconsin. Madison, Wisconsin: Department of Public Instruction, 1967.

Research in Reading[1]

Wayne Otto and Thomas C. Barrett
The University of Wisconsin
Madison, Wisconsin

Theodore L. Harris
University of Puget Sound
Tacoma, Washington

THE COMMITTEE for Research in Basic Skills has spent a number of years investigating problems of teaching and learning fundamental skills in the elementary grades. Generally, the Committee has chosen to be extensive in its approach, investigating many dimensions of a given problem. As a result, there have been a series of studies within broad problem areas.

The present investigation grew out of a recurring concern for pupils' development of flexibility in performance in the basic skill areas. In a project entitled "The Perception of Symbols in Skill Learning," Herrick, Harris, and Rarick (1961) established the importance of the development of variability rather than rigidity in motor-speed sets of handwriting. The authors pointed out that their findings might well have implications for other areas of skills development. The question of flexibility arose again in two studies (Harris and Rarick, 1957; Herrick and Otto, 1961) of the handwriting act. One implication of these studies was that efficiency in a skill such as handwriting does not seem to be associated with the development of a fixed level of pressure but rather with the ability to control pressure within given limits. Another study by Otto (1961) indicated that among good readers in the elementary grades responsiveness to auditory, visual, kinesthetic, and combined presentations of stimuli varies according to grade level and type of reinforcement. The variance, however, was not always in the direction expected, which suggests that one effect of the instructional program may be the development of perceptual sets in reading.

These studies led the Committee to the U. S. Office of Education Project No. 1755, "The Experimental Development of Variability in Rate of Reading in the Intermediate Grades" (Harris, Herrick, Macdonald, and Rarick, 1965). The results of the study showed that fourth-grade good readers who received a 2-week training program developed substantial ability to vary their speed when reading for different purposes. In the training program the children were given a) a brief introduction to three types of reading purposes – reading for the main idea, reading to find a specific fact, and reading to follow the sequence of ideas, b) short testing and training passages carefully constructed as to unity, coherence, emphasis, and appropriate content and difficulty level, c) separate selections for narrative and expository style, and d) testing to a criterion of full comprehension of each specified purpose. Specifically, the children were shown to a) possess a hitherto unreported ability to adjust significantly their reading speeds very quickly to their purposes for reading; b) alter significantly their pattern of reading speed relative to reading purposes and to maintain a significant spread in speed-purpose variability while increasing their reading speeds for each purpose approximately 40 percent, with further gains upon retention tests two weeks after training; c) adjust their reading speeds to reading purposes significantly different according to passage style and to a certain extent according to sex; d) and display the greatest improvement in reading for the main idea and specific fact. The results of Project 1755 also showed that a) children of comparable reading ability in the fifth and sixth grades have no more variability or spread in their reading speeds according to reading purposes than do fourth graders; b) children, by their own daily ratings, greatly enjoyed the 2-week training period which was largely self-instructional in nature; c) during the short training period, children became significantly more realistic in predicting their reading speeds when asked to read for sequence of ideas, but not for main idea or specific fact; and d) no untoward evidences of tension were found to be associated with reading the test passages or answering questions on them.

In keeping with its past policies, the Committee wanted to relate directly its research tasks to specific aspects of school practice or to provide the basis upon which its findings could be generalized to school programs. The experimenters felt that before schools could be expected to implement the findings of Project 1755 in their programs, they should have empirical evidence that the results of training given under the aforesaid experimental conditions would indeed transfer to more typical reading tasks faced by intermediate-grade children. The present study is, therefore, an extension of Project 1755 specifically designed to investigate the possibilities of such transfer effects.

The first part of the present study replicated the training and testing of Project 1755. The only major revision in the first part of the study was in the substitution of a second active learning group for a passive control group. Since the preceding study gave evidence that trained groups achieved greater variability than controls, it was felt to be more pertinent to test for significant differences between types of training. Therefore, half of the subjects were trained with passages containing an explicitly stated main idea and the other half worked with passages in which a main idea was implicit but not stated. Classroom materials may or may not contain explicit main idea statements so the intent was to determine whether training with either type would produce better results.

Three major hypotheses were stated:

1. Intermediate-grade children trained with passages containing either explicitly stated or implicit main ideas will differ in rate variability when reading carefully structured expository or narrative materials for three different purposes.
2. Intermediate-grade children trained with passages containing either explicitly stated or implicit main ideas will transfer their variability of reading rate to more typical materials.
3. Intermediate-grade children trained with passages containing either explicitly stated or implicit main ideas will retain their variability of reading rate with typical materials after a period of one month.

Answers to the following questions were sought:

1. Do children differ in their total reading rate when reading either narrative or expository material for three different purposes?
2. Do boys and girls in grades 4, 5, and 6 trained with passages containing specifically stated main ideas and children trained with passages in which main ideas were implicit, but not stated, differ in their rate variability when reading either expository or narrative materials for three different purposes on a training-like task immediately after training?
3. Do boys and girls in grades 4, 5, and 6 trained with passages containing either explicitly stated or implicit main ideas differ in their rate variability when reading either expository or narrative materials for three different purposes on a transfer task immediately after training?
4. Do boys and girls in grades 4, 5, and 6 trained with passages containing either explicitly stated or implicit main ideas differ in their rate variability when reading either expository or narrative materials for three different purposes on a a) training-like task immediately after training; b) transfer task immediately after training; c) transfer task one month after training?
5. Do children vary their rate when reading either expository or narrative materials for three different purposes – specific fact, sequence, or main idea on a a) training-like task immediately after training; b) transfer task immediately after training; c) transfer task one month after training?

METHOD

SUBJECTS

In the final analyses, seventy-two children from grades 4, 5, and 6 of a middle-class elementary school in Madison, Wisconsin, served as subjects. There were twenty-four subjects at each grade level, twelve boys and twelve girls. In an attempt to exclude the extremes of reading ability, subjects were chosen who had reading test scores that placed them between the fortieth and ninetieth percentiles for their grades. The STEP Reading Test was used in grade 4, the Reading Comprehension subtest of the California Achievement Test was used in grade 5, and the Iowa Reading Test was used in grade 6. School personnel corroborated the reading test scores. Table 1 shows the percentile range by grade and sex.

Within sex and grade level, equal numbers of subjects were randomly assigned to the two treatment groups. To guard against attrition due to absence, inability to cope with the task, and examiner errors, extra subjects, one boy and one girl from each grade level, were included in each treatment group. Therefore, the treatment groups comprised forty-two children, of whom six were not included in the analysis. The extra subjects remaining after attrition were randomly dropped.

EXPERIMENTAL TREATMENTS[1]

The distinction between groups was in 1) the nature of orientation to reading for the main idea and 2) the structure of materials read during the orientation and training periods. During these periods, Group I worked with materials[2] in which a main idea was explicitly stated; the task in reading for the main idea was to identify the main idea from a list of sentences taken from the passages read.

TABLE 1

PERCENTILE RANGE OF SUBJECTS' READING TEST SCORES BY GRADE AND SEX

	Grade					
	Fourth		Fifth		Sixth	
Sex	Highest Percentile	Lowest Percentile	Highest Percentile	Lowest Percentile	Highest Percentile	Lowest Percentile
Boys	88	48	90	62	87	40
Girls	90	50	90	54	89	57
Total	90	48	90	54	89	40

Group II worked with the materials from which the stated main idea had been deleted; the task in reading for the main idea involved formulating a main idea and then selecting the main idea sentence from a list identical to the one used by Group I. Both groups worked with passages of identical length and essentially the same substantive subject matter.

MATERIALS

Materials for orientation to the task, for variability training, and for immediate post training evaluation were taken, with but minor adaptations explained below, from materials constructed for Project 1755. Longer materials, more representative of the type encountered by elementary pupils, were constructed for use in the attempts to assess the transfer effects of the variability training. The development of materials was a major focus in both the present project and the one that preceded it; thus, the materials are described in considerable detail.

ORIENTATION MATERIALS

The orientation materials comprised three sets of passages, one set for each of the three purposes for reading (see description of the training program). Each passage was printed separately on charts which could be read from the location of any pupil in the testing room. A list of five names with corresponding telephone numbers was used in finding a specific fact. Reading for sequence involved one basic paragraph with three different orderings of its five sentences. For Group I orientation to reading for the main idea also employed a basic paragraph which was rewritten three times so that the topic sentence appeared in different parts of the passage. The materials were taken from those developed for Project 1755.

Because all of the Project 1755 materials included explicitly stated main ideas, the following adaptations were required for use with Group II. First, the sentence stating the main idea was deleted from the passage used to orient the subjects to reading for the main idea passage and the remaining sentences were expanded slightly to preserve a uniform length. Second, an additional paragraph was used with Group II during its orientation to the purpose of reading for the main idea. This paragraph was similar to the first one in topic and construction, but it was used to provide further clarification of the task and practice in arriving independently at a conception of the main idea. A pilot study showed that the added practice was needed.

TRAINING MATERIALS

The training materials developed in Project 1755 and used in the present study consisted of three parts; a set of directions, the passages to be read, and comprehension checks. Brief descriptions of each of the three parts of the training materials follow.

Directions

The primary purpose of the directions was to establish a definite purpose for reading either to find the Main Idea, The Sequence of Events, or a Specific Fact. Thus, a specific question was asked for each selection to be read for a Specific Fact, and when the purpose was to read for Sequence or Main Idea, the subject was told to read to find the Main Idea or to remember the order of events. The directions were designed to make the training self-instructional.

Passages

The passages were written with the following restrictions and considerations in mind. The content of the training materials was from the areas of social studies, science, or fiction because Shores' (1960) research indicates that reading speed is affected by different kinds of content.

The passages were written in narrative and expository styles because a pilot study in Project 1755 showed that expository materials were read faster than narrative, yet pupils commonly encounter both styles.

Passages were given either an inductive structure (i. e., the topic sentence came near the end of the selection) or a deductive structure (i. e., the topic sentence came near the beginning of the selection). The decision to consider structure of the material was based upon results of pilot testing in Project 1755 which indicated that children perform better on paragraphs organized deductively than inductively.

Passages were limited to six to eight sentences with total word count ranging from 66 to 112 words since the earlier pilot study showed that fourth-grade students had much difficulty with longer selections before training.

In an attempt to keep the passages within the fourth-grade readability level according to the Dale-Chall formula, the vocabulary was limited to grade 4 word lists, sentence length was limited to a seven to fourteen words range, and typical subject-verb word order was maintained within the sentences. Each passage was allowed a maximum of two examples of the syntactical complication which follows: relative and subordinate clauses, participial phrases, infinitive phrases separated from the main verb, and inverted subject-verb word order. Five teachers with experience in the intermediate grades judged the difficulty of the materials to be appropriate. As an added check, the directions for each passage encouraged the subjects to ask for vocabulary help as needed.

Each passage has a single main idea. Group I practiced with passages in which a main idea was explicitly stated; Group II practiced with passages in which the same main idea was implied but not stated. In the passages where the main idea was not stated, adjectives, adverbs, and other words which did not change the meaning of the paragraphs were inserted in the remaining sentences to yield a constant paragraph length. The nature of the daily training materials is shown in Table 2.

TABLE 2

NATURE OF THE DAILY TRAINING PROGRAM MATERIALS

Number	Selection
4	Main Idea Selections
1	Sequence Selection
1	Specific Fact Selection

Comprehension Tests

The position was taken that the reader should have access to all the passage units when selecting the appropriate main idea or determining the proper order of ideas. Thus, for the main idea and sequence tasks the comprehension tests contained a condensed or abridged restatement of each of the six to eight sentences in the passage itself and the subject either identified the main idea statement or sequentially ordered the statements given. For the specific fact tasks, a phrase containing a specific fact was taken from each sentence and the subject identified the correct one.

POST-TRAINING MATERIALS

The directions and comprehension tests used with the post-training materials were of the same type as the training materials proper and the passages were likewise adapted from paragraphs constructed for Project 1755. Six of the passages were written without a stated main idea and six were written with an explicit main idea. Each passage was sixty-six words long, including five sentences if the main idea was not stated or six sentences if it was stated.

TRANSFER MATERIALS

The central question of the study is whether variability in reading speed developed by training with carefully constructed, short model paragraphs will transfer to reading the longer, less tightly structured passages typically found in classroom material. Thus, the decision was to use the types of material actually employed in classroom instruction in the transfer task. With this decision, there was the problem of learning more about the nature and length of "typical" paragraphs in elementary level materials. A preliminary survey of tradebooks and textbooks, especially basal reading series, revealed a lack of consistency within intermediate level materials and even within individual publications.

It had at first been assumed that construction of the transfer materials would involve the use of a certain number of paragraphs, but the survey revealed that there was a surprising lack of correspondence between paragraphing and topical units. Within a single paragraph, two or three topics might be discussed; yet at another place in the same material several paragraphs might be used to cover one topic. Since one of the purposes for reading in the study involved the identification or formulation of a main idea, the experimenters decided to select passages that contained only one main idea, regardless of the total number of paragraphs involved. The survey showed also that there was little consistency among the materials with regard to the number of words used in the coverage of a single main idea. Because the experimental design of the study called for a uniform length in the transfer test materials, the experimenters arbitrarily set the length at 198 words, three times the number of words in the post-training test materials.

With consideration for all the factors involved in the looseness or tightness of the structure of materials at the intermediate grade levels, the experimenters decided that selections from the intermediate level Reader's Digest Skill Builders were reasonably representative of materials used by children in the fourth, fifthe, and sixth grades.

Passages were chosen from the Skill Builders in view of the following: each passage had to have a single, complete main idea; each passage had to be appropriately written for one of the three purposes, (i. e., if the purpose were to read for sequence, the passage must obviously contain a clearly recognizable sequential development); and all of the passages chosen had to be generally interesting to intermediate-grade children but not specifically familiar to them. The Skill Builder selections were modified only when necessary in terms of the addition or deletion of a stated main idea, and/or the addition or deletion of unimportant words in order to keep the length constant. (The authors are indebted to the editors of Reader's Digest for permission to use their materials and for their examination and approval of all adaptations made in the materials.) Half of the selections chosen were written in a primarily narrative style and half were expository in style.

Equivalent sets of materials were necessary for the immediate and delayed transfer testing. "Pairs" of materials were developed by drawing two selections from the same Skill Builder article. Since these articles were very short, it was felt that vocabulary, style, and other factors affecting readability were likely to be reasonably equivalent. The selections in each pair were assigned randomly to the immediate or delayed transfer testing materials.

Table 3 is a breakdown of the total number of selections used for the transfer study.

TRAINING AND TESTING PROCEDURE

In keeping with the framework established by Project 1755, the research program was divided into two basic parts, training and testing, which were then subdivided with time designated for each part as shown in Table 4.

Training Program

Training Examiners. A week before the training program began, the adult assistants met with the graduate research staff for orientation to methods and procedures to be used. A brief history of the research project was given, and each assistant received a copy of the complete set of directions for each day of training. Objectivity in working with the subjects was emphasized.

Training Subjects. Days 1-3 of the training program served as an Orientation Period to prepare subjects to read for the three different purposes and to familiarize them with the materials and procedures. The subjects read three practice passages each day. The orientation sequence is shown in Table 5.

The difference between groups in the Orientation Period was in presentation of reading for the main idea. Group I practiced with stated main ideas in their paragraphs while Group II practiced with paragraphs with implicit main ideas. This distinction was maintained throughout the training program.

On days 4-10 the subjects practiced for 30-45 minute periods each day, during which they read six selections daily. The selections were randomly ordered for the total group of eighty-four subjects every day.

In addition to reading passages, subjects also practiced predicting their reading speed for each selection. This was followed by recording actual speed on their individual graphs and on Time Record Sheets. It was felt that this would help each subject evaluate his own progress; the graph would enable him to visualize the differences in reading rate for each of the three purposes for reading.

TABLE 3

NATURE OF THE SELECTIONS USED IN THE IMMEDIATE AND DELAYED TRANSFER SESSIONS

Selections for Immediate Transfer		Selections for Delayed Transfer	
Stated Main Idea			
2*	Main Idea	2	Main Idea
2	Sequence	2	Sequence
2	Specific Fact	2	Specific Fact
Unstated Main Idea			
2	Main Idea	2	Main Idea
2	Sequence	2	Sequence
2	Specific Fact	2	Specific Fact

*There was one narrative and one expository selection in each pair.

On each of the seven practice days, this procedure was followed:

1. Each subject received his set of self-instructional materials and read the directions to establish a purpose for reading.
2. Each subject predicted his reading speed on his graph.
3. Each subject timed his reading of the selection with a stop watch and recorded actual reading time on Time Record Sheets.
4. An answer sheet accompanying the selection was then completed.
5. Subjects were expected to read to full comprehension or 100 percent accuracy in responding to the comprehension test items. Due to the nature of the tests, five trials was the maximum. The assistants checked individual responses after each trial.
6. Raw time scores were translated into words per minute scores by the subjects, who were supplied with conversion tables.
7. Subjects then plotted reading speed on their graphs to show a comparison of predicted and actual reading speed. The graphing was intended to serve as a motivational device.

ADMINISTRATION OF THE TRAINING PROGRAM

To obtain manageable groups for the training sessions, Groups I and II were each divided into two training groups of twenty-one subjects. Three groups met separately in the testing room during the morning, and one group met in the afternoon.

Four research assistants conducted the Training Program. Since each had been involved in working on various phases of the training materials and training procedure, it was felt that they would be highly consistent in their administration of the four training groups. Four or five adult assistants were

TABLE 4

SEQUENCE OF TRAINING AND TESTING

Training Program		Testing Program		
Orientation Period	Training Period	Post-Training Test Period	Immediate Transfer Test Period	Delayed Transfer Test Period
Days 1-3	Days 4-10	Day 11	Days 12-13	Days 14-15 (one month following Days 12-13

TABLE 5

ORIENTATION SEQUENCE

Day 1	Day 2	Day 3
Introduction to (1) three purposes for reading (2) use of stop watches (3) use of Time Record Sheets	Introduction to (1) Conversion tables (2) Graphs	Introduction to (1) predicting reading speed

available at each session to supervise the subjects' work. The resulting ratio of one adult for every four or five subjects insured adequate supervision of the work and immediate feedback on the comprehension tests.

The following administrative details were patterned after Project 1755:

1. Training groups were no larger than twenty-one subjects.
2. A large, well-ventilated, well-lighted, quiet room was used.
3. Individual work space was available for each pupil.
4. All necessary supplies were provided by the researchers.
5. One adult was assigned to work with every four or five subjects.
6. Instructors and assistants were trained in the testing procedures and to anticipate possible problems.
7. Detailed printed instructions were provided for instructors and assistants.

TESTING PROGRAM

The procedure followed during the three phases of the Testing Program (see Table 4) deviated from the Training Program in these ways: (1) For the Post-Training Test, day 11, 12 randomly ordered, sixty-six word selections were read instead of the usual six selections of varying length. Four randomly ordered selections were read for each of the three purposes. The total reading time in seconds for each purpose was recorded. The pupils did not use their graphs. (2) The Immediate and Delayed Transfer selections were 198 words in length. They are described in detail in the materials section of this paper. All subjects read twelve randomly ordered selections, six with and six without an explicitly stated main idea, for Immediate Transfer on days 12 and 13. One month later, on days 14 and 15, they read twelve selections comprising the second half of the set of transfer materials. During the two transfer testing periods, the assistants recorded the subjects' total reading time in seconds when reading for each purpose. No graphs or conversion tables were used.

The assumption was that in order to assess the effects of the stated and implied main idea treatment conditions it was necessary to test the subjects on materials with and without explicitly stated main ideas. A reasonable expectation would be that subjects trained with stated main ideas would do best on test materials which contained stated main ideas and those trained with materials having implicit main ideas would do best on test materials with implicit main ideas. To assess the overall effect of the treatment conditions, the experimenters were concerned with the best performance on both types of materials. In order to get a score which indicated this total performance, regardless of type of material, the mean score of the subjects' performance on narrative and expository material was found, thus reducing the number of scores for each subject to six.

DESIGN AND ANALYSIS

In order to determine the variability of reading rate for the three different purposes – specific fact, sequence, and main idea – a variability score was calculated for each subject. The style of the material was considered in that two rate variability scores were calculated for each subject – one for narrative and one for expository materials. The variability scores were calculated according to the standard variance formula:

$$S^2 = \sum_{i=1}^{3} \frac{(Xi - \overline{X})^2}{n - 1} \ .$$

Xi was the sum of individual scores where i varied across the three purposes; $\overline{X}$ was the mean of the three purpose scores, and n = 3. Thus, the variance scores took into account the subjects' reading rate for one purpose in relation to all purposes (mean rate). There were three different designs and consequently three different types of analysis.

One was a subject × purpose one way analysis of variance design. The dependent variable for the analysis was the proportion of each individual's reading rate for each single purpose to his total reading rate for all three purposes.

Due to the fact that each subject read for three purposes, a repeated measures situation existed. Greenhouse and Geisser (1959) have pointed out that for the usually computed F ratios of mean squares in the regular analysis of variance to be exactly distributed as the F distribution, it is necessary that the columns or tests in addition to being normally distributed, having equal variances, should be mutually independent. Since the measures were made on the same subjects, the independence assumption could not be met.

Although one approach to this problem would be to use a multivariate analysis of variance design, a simpler approach is to follow the procedure of adjusting the degrees of freedom which results in conservative F ratios which approximate the F distribution when the independence assumption is not met (Greenhouse and Geisser, 1959). The adjusted degrees of freedom are in this case 1 and 71 respectively. Six analyses were also done using this design and procedure.

The second was a 3 × 2 × 2 completely crossed analysis of variance design with three grades, two sexes, and two training methods and with six replicates per cell. The model was assumed to be fixed. The reading variability scores were transformed by using the natural log to base e. Thus, the scores (y) used in the analysis were:

$$y = \text{nat log}_e \sum_{i=1}^{3} \frac{(Xi - \overline{X})^2}{n - 1} \ .$$

After the scores were transformed, an analysis of variance was done on the means of the variances according to the procedure outlined by Scheffé (1960) for the comparison of variances. Six analyses were done according to these design and analysis procedures.

The third design was a 3 × 2 × 2 × 3 analysis of variance with three grades, two sexes, two training procedures, and three repeated measurements or tests – a training test immediately after training, a transfer task immediately after training, and a delayed transfer task one month after training. The dependent variable for the analyses using this design was rate variability. Two analyses were done using this procedure. Conservative F tests using the Greenhouse and Geisser (1959) procedure were also used in this analysis. The adjusted degrees of freedom were 1 and 69 respectively. For all F ratios found to be significant post hoc comparisons among the means were done following the procedures of Scheffé (1953).

RESULTS AND DISCUSSION

The subjects consistently varied their rate when reading either narrative or expository material for three different purposes. The analyses summarized in Table 6 also show that there was a difference in rate variability on the immediate post training test, the transfer task, and the transfer retention task.

Mean proportions of reading time are shown in Table 7. The trend there is clear and, with but one exception, consistent: the children read most rapidly to find specific facts; they read less rapidly for main ideas, and they read most slowly for sequence. The Scheffé post hoc test for multiple comparisons reveals significant differences between all means except the Post-Narrative, Main Idea, and Specific Fact. This is in line with the logical expectation that, with training, children would be able to adapt their rate of reading according to purpose; that they would be best able to pick out specific facts very quickly; and that the task of reading to arrange facts in sequence would be most time consuming because the subjects would need to deal simultaneously with several facts.

Materials were written in narrative and expository styles because there was evidence that style affects rate. When the means in Table 7 were collapsed across style, there was virtually no difference between the narrative and expository means (Expository = .340; Narrative = .333). Yet, a comparison that takes purpose into account, as shown in Table 8, reveals that style interacts with purpose; expository materials are read most rapidly for the main idea, but narrative materials are read most rapidly for sequence and for specific fact. By way of post hoc interpretation, it would appear that the expository arrangement makes for more clearly identifiable main ideas. Likewise, the narrative arrangement of materials would tend to have a sequential arrangement built in and, therefore, be more easily identifiable. An explanation of the more rapid reading of narrative materials for specific fact is more elusive; but it might be argued that a narrative-sequential arrangement lends clues that make for rapid location of facts.

TABLE 6

ONE WAY ANALYSES OF VARIANCES FOR TOTAL READING RATE FOR THREE DIFFERENT PURPOSES WITH EXPOSITORY AND NARRATIVE MATERIALS ON POST, IMMEDIATE, AND DELAYED TESTS

Description	df	MS	F
Purposes on Expository Posttest	2 (1)	.40	40*
Error	213 (71)	.01	
Purposes on Narrative Posttest	2 (1)	.72	72*
Error	213 (71)	.01	
Purposes on Expository Transfer Test	2 (1)	.97	97*
Error	213 (71)	.01	
Purposes on Narrative Transfer Test	2 (1)	.92	92*
Error	213 (71)	.01	
Purposes on Expository Retention Test	2 (1)	.84	84*
Error	213 (71)	.01	
Purposes on Narrative Retention Test	2 (1)	.36	36*
Error	213 (71)	.01	

* $p < .01$

TABLE 7

MEAN PROPORTION OF READING TIME FOR TOTAL COMPREHENSION BY STYLE OF MATERIAL, TIME OF TEST AND PURPOSE

Test and Style	Purpose		
	Main Idea	Sequence	Fact
Post-Expository	.331	.409	.261
Post-Narrative	.389	.393	.218
Immediate Transfer-Expository	.359	.435	.207
Immediate Transfer-Narrative	.348	.438	.214
Delayed Transfer-Expository	.325	.446	.229
Delayed Transfer-Narrative	.382	.366	.252

TABLE 8

MEAN PROPORTION OF READING TIME BY STYLE AND PURPOSE

Style of Material	Purpose		
	Main Idea	Sequence	Fact
Expository	.338	.430	.252
Narrative	.373	.399	.228

TABLE 9

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR EXPOSITORY MATERIALS AND POST PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	5.98	2.89
Sex	1	1.00	< 1
Treatment	1	.55	< 1
G × S	2	2.99	1.46
G × T	2	4.30	2.10
S × T	1	1.14	< 1
G × S × T	2	.84	< 1
Error	60	2.05	

TABLE 10

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR NARRATIVE MATERIALS AND POST PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	3.76	1.52
Sex	1	2.07	< 1
Treatment	1	4.11	1.66
G × S	2	.64	< 1
G × T	2	.23	< 1
S × T	1	1.17	< 1
G × S × T	2	1.24	< 1
Error	60	2.48	

The type of materials used during the training sessions did not make a difference in the children's subsequent variability rate when reading for three different purposes (see Tables 9-14). Children trained with passages containing explicitly stated main ideas did not differ in rate variability from children trained with materials with implicit but not specifically stated main ideas. The types of training materials are subsequently referred to as stated and unstated, respectively. The implication is that materials varied on this aspect do not affect children's ability to vary their reading rate. Thus, either type of training materials should produce similar results if reading variability across purposes is the objective of the training.

The use of different types of training materials did not have any effect on the transfer task. When children were given an extended passage which was more closely related to the actual school reading task, children trained with stated materials did not differ in their rate variability from children trained with unstated materials.

The same was true for the transfer retention task. When children were tested for retention on the transfer task, no differences in rate variability were found between the two groups trained with different materials.

However, one might speculate that if the approach to instruction had been varied for each type of materials, the results might have been different. For example, if the children who worked with unstated materials had been trained explicitly in devising their own main idea statements, they might have done better than the other group when all subjects were tested on both stated and unstated materials. Such speculation seems to suggest a possible next step for research that focuses upon training programs for the development of rate variability.

Whether the children were reading expository or narrative materials, the results were the same. Consistently on the post, immediate transfer, and delayed transfer tests no differences were found between the group trained with stated main idea materials and the group trained with unstated main idea materials.

TABLE 11

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR EXPOSITORY MATERIALS AND IMMEDIATE PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	4.94	2.66
Sex	1	4.22	2.27
Treatment	1	.34	< 1
G × S	2	2.35	1.26
G × T	2	2.40	1.29
S × T	1	4.43	2.38
G × S × T	2	2.40	1.29
Error	60	1.86	

TABLE 12

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR NARRATIVE MATERIALS AND IMMEDIATE PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	.71	< 1
Sex	1	5.33	1.60
Treatment	1	.00	< 1
G × S	2	1.51	< 1
G × T	2	5.10	1.53
S × T	1	.32	< 1
G × S × T	2	.03	< 1
Error	60	3.33	

Sex did not make a difference in the rate variability of children trained with either of the two types of materials. This result was consistent. Boys did not differ from girls in their ability to vary their rate from purpose to purpose immediately after training, on the transfer task or on the transfer retention task.

The same results held across grade level. There was no difference among fourth, fifth, and sixth graders in reading rate variability from purpose to purpose on each separate testing occasion. However, the trends indicated by the means (Table 15) are consistent. Fourth graders consistently showed more variability of reading rate than did fifth graders, and fifth graders consistently showed more variability of reading rate than did the sixth graders. This could be due to the younger children's more flexible mental set, which could make them much more susceptible to training.

The absence of grade by treatment interaction indicates that the two different training procedures did not have different effects across grade level. This might not have been so if materials had been prepared for each specific grade level. Further investigation that focuses upon possible developmental trends might be worthwhile.

The total reading variability across occasions (tests) on expository materials was found to differ by grade (Table 16). No grade differences were apparent, however, on the total reading variability across occasions on narrative materials (Table 17). Mean rate variability by grade and type of material is shown in Table 18. Scheffé post hoc tests for multiple comparisons revealed that the only significant difference is between the mean scores of fourth and sixth graders when reading expository material. This supports the trend found with individual tests. By way of interpretation, the suggestion is that older pupils may have a fixed set for approaching the reading task; whereas, younger pupils may be more readily susceptible to rate variability training. If this is so, the implication is that such training ought to be given early in the intermediate grades.

TABLE 13

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR EXPOSITORY MATERIALS AND DELAYED PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	2.97	1.40
Sex	1	.16	< 1
Treatment	1	.27	< 1
G × S	2	.23	< 1
G × T	2	.31	< 1
S × T	1	.71	< 1
G × S × T	2	1.93	< 1
Error	60	2.12	

TABLE 14

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR NARRATIVE MATERIALS AND DELAYED PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	5.80	2.10
Sex	1	1.78	< 1
Treatment	1	.15	< 1
G × S	2	2.95	1.07
G × T	2	2.45	< 1
S × T	1	7.36	2.67
G × S × T	2	2.44	< 1
Error	60	2.76	

TABLE 15

MEANS FOR EACH GRADE ON EXPOSITORY AND NARRATIVE MATERIALS ON THREE DIFFERENT TEST OCCASIONS

Type	Occasion	Grade 4	Grade 5	Grade 6
Expository	Post	3.07	2.66	2.08
	Immediate	6.29	6.20	5.46
	Delayed	6.40	5.89	5.72
Narrative	Post	3.45	3.35	2.72
	Immediate	5.66	5.92	5.59
	Delayed	5.80	5.61	4.87

TABLE 11

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR EXPOSITORY MATERIALS AND IMMEDIATE PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	4.94	2.66
Sex	1	4.22	2.27
Treatment	1	.34	< 1
G × S	2	2.35	1.26
G × T	2	2.40	1.29
S × T	1	4.43	2.38
G × S × T	2	2.40	1.29
Error	60	1.86	

TABLE 12

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR NARRATIVE MATERIALS AND IMMEDIATE PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	.71	< 1
Sex	1	5.33	1.60
Treatment	1	.00	< 1
G × S	2	1.51	< 1
G × T	2	5.10	1.53
S × T	1	.32	< 1
G × S × T	2	.03	< 1
Error	60	3.33	

Sex did not make a difference in the rate variability of children trained with either of the two types of materials. This result was consistent. Boys did not differ from girls in their ability to vary their rate from purpose to purpose immediately after training, on the transfer task or on the transfer retention task.

The same results held across grade level. There was no difference among fourth, fifth, and sixth graders in reading rate variability from purpose to purpose on each separate testing occasion. However, the trends indicated by the means (Table 15) are consistent. Fourth graders consistently showed more variability of reading rate than did fifth graders, and fifth graders consistently showed more variability of reading rate than did the sixth graders. This could be due to the younger children's more flexible mental set, which could make them much more susceptible to training.

The absence of grade by treatment interaction indicates that the two different training procedures did not have different effects across grade level. This might not have been so if materials had been prepared for each specific grade level. Further investigation that focuses upon possible developmental trends might be worthwhile.

The total reading variability across occasions (tests) on expository materials was found to differ by grade (Table 16). No grade differences were apparent, however, on the total reading variability across occasions on narrative materials (Table 17). Mean rate variability by grade and type of material is shown in Table 18. Scheffé post hoc tests for multiple comparisons revealed that the only significant difference is between the mean scores of fourth and sixth graders when reading expository material. This supports the trend found with individual tests. By way of interpretation, the suggestion is that older pupils may have a fixed set for approaching the reading task; whereas, younger pupils may be more readily susceptible to rate variability training. If this is so, the implication is that such training ought to be given early in the intermediate grades.

TABLE 13

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR EXPOSITORY MATERIALS AND DELAYED PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	2.97	1.40
Sex	1	.16	< 1
Treatment	1	.27	< 1
G × S	2	.23	< 1
G × T	2	.31	< 1
S × T	1	.71	< 1
G × S × T	2	1.93	< 1
Error	60	2.12	

TABLE 14

ANALYSIS OF VARIANCE OF GRADE, SEX, AND TREATMENT FOR NARRATIVE MATERIALS AND DELAYED PERIOD READING VARIABILITY FOR THREE PURPOSES

Source of Variation	df	MS	F
Grade	2	5.80	2.10
Sex	1	1.78	< 1
Treatment	1	.15	< 1
G × S	2	2.95	1.07
G × T	2	2.45	< 1
S × T	1	7.36	2.67
G × S × T	2	2.44	< 1
Error	60	2.76	

TABLE 15

MEANS FOR EACH GRADE ON EXPOSITORY AND NARRATIVE MATERIALS ON THREE DIFFERENT TEST OCCASIONS

Type	Occasion	Grade 4	Grade 5	Grade 6
Expository	Post	3.07	2.66	2.08
	Immediate	6.29	6.20	5.46
	Delayed	6.40	5.89	5.72
Narrative	Post	3.45	3.35	2.72
	Immediate	5.66	5.92	5.59
	Delayed	5.80	5.61	4.87

TABLE 16

ANALYSIS OF VARIANCE OF GRADE, SEX, TRAINING AND TEST OCCASION FOR READING VARIABILITY WITH THREE PURPOSES AND EXPOSITORY MATERIALS

Sources of Variation	df	MS	F
Training	1	.10	.05
Sex	1	2.83	1.35
Grade	2 (1)	13.57	6.49*
Occasions	2 (1)	36.23	17.32*
T × S	1	.88	.42
T × G	2	3.04	1.45
T × O	2	.37	.18
S × G	2	2.26	1.08
S × O	2	1.57	.75
G × O	4	.74	.35
T × S × G	2	2.51	1.20
T × S × O	2	3.02	1.44
T × G × O	4	1.62	.78
S × G × O	4	1.49	.71
T × S × G × O	4	1.53	.73
Within (error)	180 (69)	2.09	

* $p < .05$

TABLE 17

ANALYSIS OF VARIANCE OF GRADE, SEX, TRAINING AND TEST OCCASION FOR READING VARIABILITY WITH THREE PURPOSES AND NARRATIVE MATERIALS

Sources of Variation	df	MS	F
Training	1	.86	.30
Sex	1	8.61	3.01
Grade	2	7.87	2.75
Occasions	2	2.55	.89
T × S	1	1.61	.56
T × G	2	1.28	.44
T × O	2	1.70	.60
S × G	2	1.24	.43
S × O	2	.28	.10
G × O	4	.20	.42
T × S × G	2	2.34	.82
T × S × O	2	3.62	1.26
T × G × O	4	3.25	1.13
S × G × O	4	1.93	.67
T × S × G × O	4	.69	.24
Within (error)	180	2.86	

The analyses summarized in Tables 16 and 17 also reveal a difference in rate variability across tests – post-training, immediate transfer, and delayed transfer – with expository materials but not with narrative material. Mean rate variability by test and type of material is shown in Table 19. Scheffé post hoc tests show a significant difference between the post-training and the immediate and delayed transfer tests with expository material. The suggestion is that not only does the ability to adjust rate to purpose transfer from training passages to more naturalistic materials, but also there is an increase in rate variability. The increase might be expected on the basis that the materials in the

TABLE 18

MEAN RATE VARIABILITY ON ALL TESTS BY GRADE LEVEL AND TYPE OF MATERIAL

Type	Grade 4	Grade 5	Grade 6
Expository	5.98	5.65	5.12
Narrative	5.70	5.93	5.12

transfer test situations were more nearly like classroom materials and less rigidly structured; therefore, the subjects were better able to use their rate variability skill.

It must be noted that the main effects of grade and test were significant only with expository materials. Yet, the trends were generally similar. The structure of the expository material apparently was more conducive to rate variability than that of narrative material. Whether this is generally so, of course, remains to be seen.

There was no difference in the total reading variability of boys and girls. Boys and girls thus learn the same amount during the training situations. No differences were found on any of the interaction factors, which indicates that all factors were consistent in combination.

The type of training materials used did not result in a difference on the total reading variability across occasions regardless of whether the children were reading expository or narrative material. This again emphasizes the similar effect of either of the two types of training materials. It may also indicate that in the future this type of research ought to be concerned with combinations of materials and instructional procedure in order to maximize the effect of training.

TABLE 19

MEAN RATE VARIABILITY FOR EACH TEST BY TYPE OF MATERIAL

Type	Test: Post	Test: Immediate	Test: Delayed
Expository	4.77	5.99	6.01
Narrative	5.37	5.73	5.43

FOOTNOTES

1. The research reported herein was supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education and Welfare. This paper first appeared in manuscript form entitled "Transfer Effects of Training Intermediate Grade Pupils to Adjust Reading Speed to Reading Purpose," Cooperative Research Project No. 3137, Committee for Research in Basic Skills, The University of Wisconsin, 1966.

2. An appendix containing the directions and materials used can be obtained from the authors on request.

REFERENCES

Greenhouse, S. W., and Geisser, S. On methods in the analysis of profile data. Psychometrika, 1959, 24, 95-112.

Harris, T. L., Herrick, V. E., Macdonald, J. B., and Rarick, G. L. Experimental development of variability in reading rate in grades 4, 5, and 6. Cooperative Research Project No. 1755, Committee for Research in Basic Skills. Madison, Wisconsin: School of Education, University of Wisconsin, 1965.

Harris, T. L., and Rarick, G. L. The problem of pressure in handwriting. Journal of Experimental Education, 1957, 26, 115-178.

Herrick, V. E., Harris, T. L., and Rarick, G. L. Perception of symbols in skill learning by mentally retarded, gifted, and normal children. Committee for Research in Basic Skills. Madison, Wisconsin: School of Education, University of Wisconsin, 1961.

Herrick, V. E., and Otto, W. Pressure on point and barrel of a writing instrument. Journal of Experimental Education, 1961, 30, 215-230.

Otto, W. The acquisition and retention of paired associates by good, average, and poor readers. Journal of Educational Psychology, 1961, 52, 241-248.

Scheffé, H. A. A method for judging all possible contrasts in the analysis of variance. Biometrika, 1953, 40, 87-104.

Scheffé, H. The analysis of variance. New York: Wiley, 1959.

Shores, H. J. Reading science materials for two distinct purposes. Elementary English, 1960, 37, 546-552, 565.

Strategies for Concept Attainment in Mathematics

Myron F. Rosskopf
Teachers College, Columbia University
New York, New York

IT WOULD be a mistake for the writer, being trained in mathematics rather than psychology, to do basic research in conceptual learning. On the other hand, it is quite appropriate for him to acquaint himself with the results of experimentation by psychologists and to manipulate these results in different ways to see how they apply in mathematics instruction. In the latter sense, one might say the paper has an operational flavor. That is, it discusses ways classroom teachers might attack instruction in a specific concept of mathematics and means writers of classroom materials might use to insure that provision had been made for learning experiences that would lead to attainment of a specific concept.

A teacher who faces a mathematics class five days each week often finds psychological research of very little help in solving his instructional problems. The research, for obvious reasons, must start with those situations that present knowledge can handle. A research psychologist legitimately pursues his interest in pure psychology; he may have fleeting thoughts about applications of his research, but the excitement of carrying on investigations in his particular field of interest leads him to put these thoughts aside – temporarily, perhaps.

Today, however, there exists sufficient evidence on human learning as well as numbers of well-trained mathematics teachers so that applications can be made in classrooms. It is possible today to have classroom experimentation that makes use of, or manipulates, basic psychological research in human learning. Emerging among mathematical educators and psychologists are plans for studies that will apply in a subject matter field the results obtained from the intensive study of conceptual learning of recent years (that is, since World War II). Every discipline has its own subtleties. For this reason, success of such experimentation will depend upon facility of communication between psychologists and representatives of other disciplines.

DEFINITIONS

CONCEPT

The word "concept" seems to have as many definitions as there are writers on conceptual learning. Recently, a definition appeared that has the virtue of efforts made to relate it to existing definitions. According to this recent statement,

> Anything that is a concept has the following attributes:
> I. Psychological meaningfulness
> II. Intrinsic, functional, or formal properties
> III. Abstractness
> IV. Inclusiveness
> V. Generality
> VI. Structure
> VII. Function (Klausmeier, Davis, Ramsay, Fredrick, and Davies, 1965, p. 3).

A possible criticism is that the definition introduces words which in turn need to be defined. True, but it turns out to be easier to define (or describe) these new words than it was to define "concept."

Two aspects of the definition that make it useful are its open-endedness – by means of item V, Generality – and its own inclusiveness. Simple concepts as well as very complex concepts are possible of characterization by means of the definition.

The writer has been using the following definition of a concept, probably because it lent itself so well to explanations of mathematical concepts. A concept is "a way of grouping an array of objects or events in terms of those characteristics that distinguish this array from other objects or events in the universe" (Bruner, Goodnow, and Austin, 1956, p. 275). This suffices for a first level concept, but to obtain the complexity that we know, at least intuitively, some concepts have, we need to add to the definition cited; and including "the network of inferences that are or may be set into play by an act of categorization" (Bruner, et al, 1956, p. 244).

We shall not stop to relate these two definitions. Interested readers can go to the references cited and see that the two do indeed contain the same elements. Instead we shall consider an example of a concept in mathematics and show in what way it satisfies the conditions of the first definition stated. Consider, for example, the set of all counting numbers, or as they are often called the set of all natural numbers,

$[1, 2, 3, \ldots, n, n+1, \ldots]$.

One can discriminate certain subsets of this set. There is the set of all even natural numbers,

$[2, 4, 6, \ldots, 2n, \ldots]$;

and the set of all odd natural numbers,

$[1, 3, 5, \ldots, 2n+1, \ldots]$.

For our example of a mathematical concept, however, we shall consider that of the set of all prime (natural) numbers,

$[2, 3, 5, 7, 11, 13, \ldots]$.

When a person can distinguish between prime numbers and numbers that are not prime (that is, that are composite), then he has a concept of prime numbers. He can demonstrate his possession of the concept by filling in the next few prime numbers, or by devising a test for deciding whether a number presented to him is prime, or not prime. He has grouped the array of objects, the set of natural numbers, so that he can tell the difference between a prime number and a composite number. He has used the characteristics that distinguish between prime numbers and other members of the set of natural numbers. Really, the person has formed two categories of natural numbers; those that are prime form one category; and those that are composite form the other category.

At this stage of his concept of a prime number, a person recognizes the formal properties. He certainly is working at an abstract level, since he is working with numbers, and at a level of generality, since many subsidiary concepts such as factorization, multiplication, and so on are involved. From the point of view of structure, the person probably has used a conjunction of related concepts to arrive at a concept of prime numbers.

But we recognize more in a concept than just the forming of categories according to some discriminating characteristic or characteristics. Part of the concept of prime numbers is realization that their number is infinite, that there is just one even prime, that no one has devised a formula for prime numbers, and that there are many such open questions related to this interesting subset of the set of natural numbers. In other words, we are willing to accept as part of the concept these inferences drawn from the categorization.

It seems that such associations with the set of prime numbers would qualify it for the attribute of psychological meaningfulness. There remain the attributes of inclusiveness and function. Now, inclusiveness is so much a part of specifying a set in mathematics that as soon as one is able (that is, correctly in the sense of mathematical criteria) to talk about the set of prime numbers, the attribute of inclusiveness follows. Prime numbers as mediators occur in prime factorization, calculation of a least common multiple, and the prime factorization theorem of arithmetic. Hence, the concept has the function attribute also.

All right. It is possible to relate what mathematicians call a concept to what psychologists call a concept. But a definition, if it really is a definition must partition the universe of discourse into those things that are concepts and those things that are not concepts. A "definition" that is all inclusive might just as well not be stated at all. In short, it must be possible to point to something in mathematics that is not a concept. But that is easy; any addition fact, such as $5 + 4 = 9$, is not a concept.

CATEGORIZING OR CONCEPTUALIZING

The word "category" was used in the preceding discussion. What is involved in categorizing? How is it related to conceptualizing? More important, how is categorization related to the use of conceptualization in applications, or problem solving, in mathematics? First, let us analyze another example of a mathematical concept and, second, derive from the discussion of the example what seem to be characteristics of categorizing.

Consider the concept of a rational number. Its formation starts very early in the home experiences of a preschool child. One-half and one-fourth are known for names of parts of things; one-half of an apple or an orange or a glass of juice or a candy bar. During early schooling, the concept is expanded to include written symbols for one-half and one-fourth: 1/2 and 1/4; 1/3 also is added to the list. However, the concept is still on the level of these symbols being names for parts of things. There seems little evidence that a child thinks of 1/2, 1/4, 1/3 as exemplars of a new sort of number. The teacher may call these fractions names for numbers, of course, but still the concept, the way of categorizing these numbers, of telling how they differ from other numbers, is in terms of a way of expressing them; to wit, a numerator, a horizontal bar, and a denominator.

Not until quite late does the instruction reach the stage of giving a new name, rational numbers, to the set of numbers that can be expressed as the quotient of two integers, the denominator not zero. Now the concept has formal properties; it has been formalized, one might say, and it will be even more formal (that is, abstract) when a student recognizes rational numbers as ordered pairs of integers in which the second member is not zero.

At each level of formality, there are cues that a student looks for: 1) one of two congruent parts of something, such as an apple; 2) a numerator, a horizontal bar, and a denominator; 3) a quotient of two integers; and 4) an ordered pair of integers. These cues help a student put a number he encounters into its proper category, to align it with the concept he has formed of the set to which that number belongs. Notice, too, that the invention of ways of grouping predictive characteristics is involved; for example, the numerator, horizontal bar, and denominator. The test of the invention is its usefulness. Does it effectively categorize numbers?

Categorizing has for its basis defining characteristics. By grouping together things that satisfy the defining characteristics of a class, an individual a) reduces the complexity of his environment; b) identifies the objects in a particular universe of discourse; c) reduces the necessity for constant learning; d) gives direction to activity; e) orders and relates classes (Bruner, et al., 1956, pp. 12-13).

What are the implications for problem solving in mathematics? If a student does not classify problems he meets, then he must treat each problem as an individual. The cognitive strain in such a procedure is enormous. The student must, perforce, for purposes of economy of effort look for distinguishing characteristics or discriminable attributes of problems in order to reduce the strain. That is, he must use cues, put the problem in a pre-decided-upon category, and bring into play concepts that in the past have been effective in solving problems of this particular category. It would seem that a categorization of a problem is the first step toward its solution. Only after a problem has been put in a certain category can one proceed with plans for solving it.

Of course the first plan chosen may fail. If it does, a person has to devise variations on the first plan. Perhaps he failed to discriminate carefully enough when he categorized the problem. In other words, taking a first step toward its solution, leads him to recategorize the problem. The important thing is that he must realize that the initially chosen method of solution fails and then search for a new method. He must shift his plans, and this requires strategy.

STRATEGIES IN CONCEPTUALIZATION

STRATEGY

A strategy is a carefully planned calculation and coordination of the specific ends and means necessary to achieve a goal. That is, as in the following:

> A strategy refers to a pattern of decisions in the acquisition, retention, and utilization of information that serves to meet certain objectives, i.e., to insure certain forms of outcome and to insure against certain others. Among the objectives of a strategy are the following:
>
> a. To insure that the concept will be attained after the minimum number of encounters with relevant instances.
>
> b. To insure that a concept will be attained with certainty, regardless of the number of instances one must test en route to attainment.
>
> c. To minimize the amount of strain on inference and memory capacity while at the same time insuring that a concept will be attained.
>
> d. To minimize the number of wrong categorizations prior to attaining a concept. (Bruner et al., 1956, p. 54.)

The preceding definition of a strategy and a description of the aims of a strategy is very similar to the mathematical definition of a strategy that is associated with the theory of games. However, the theory of games strategy consists of a set of decisions made in advance. Once the decisions are made, no change can be made while "play" is going on. In contrast, a problem-solving strategy, for example, may be changed in the midst of a subject's efforts; such changes are dictated, or should be dictated, as the subject's knowledge of the problem grows through his efforts to find a method of solution.

TYPES OF STRATEGIES

"Virtually all the effective strategies for attaining concepts depend upon the use of some sort of initial focus," (Bruner, et al., 1956, p. 63). How does the subject perceive the situation? What aspects of it holds his attention? How does he define the task to himself? If the task is a problem situation, what data seem to be most relevant? How does he fit this problem into his concept pattern of all types of problems with which he has had experience? After making an initial focus with respect to the problem, a subject still faces the necessity for decisions concerning a method of attack. What sort of strategy should he decide upon? According to Bruner and his associates there are four ideal strategies: a) simultaneous scanning; b) successive scanning; c) conservative focusing; and d) focus gambling (Bruner, et al., 1956, pp. 83-90).

a) Simultaneous scanning. The simultaneous scanner forms many possible hypotheses to follow to a solution and keeps all of these hypotheses in mind. The strain on memory is great. Initially, he chooses one of these hypotheses and tests it to see if it is promising and at the same time to see which of the other hypotheses it eliminates. A subject using this strategy has to keep all of the hypotheses in mind; he has to deduce which are eliminated, and as he proceeds to test a remaining hypothesis, he must remember which hypotheses were eliminated by his earlier testing.

The critical point in simultaneous scanning depends upon a subject's ability to deduce all the information with respect to which hypotheses should be discarded from the testing of an initial hypothesis. If he can do this, then he can proceed with maximum information to test a remaining hypothesis. However, memory strain and the factor of time usually prevent him from obtaining maximum information. Suppose the subject makes a wrong deduction? Does he have control over this risk? He does not; he cannot himself either increase or decrease the risk.

b) Successive scanning. Essentially, this strategy consists of testing one hypothesis at a time. A hypothesis is tested until a contradiction is reached, or until it is clear to the subject that the desired information is not forthcoming. Then the hypothesis is discarded, and a second is chosen for testing. Notice that a hypothesis is tested independent of the others.

The gain in use of successive scanning is twofold: 1) it lightens memory load and 2) frees the subject from the necessity for making many deductions concerning the other hypotheses. Two disadvantages are that some of the hypotheses may overlap, leading to repetition of work already done, and that there is no control over risk. A subject cannot take greater gambles or lesser gambles. This is a rather leisurely strategy and discounts heavily any time factor.

c) Conservative focusing. First of all, a subject who is a conservative focuser studies the situation to see if there is one key idea. If there is, it is focused upon, and all hypotheses are tested against it. Suppose he finds a hypothesis that looks promising. That is, a hypothesis that impinges on the key idea. Now, he chooses those hypotheses that contain the same part that the promising hypothesis does. Then, one by one, he adds parts of these hypotheses and tests the results. The "conservative" enters this strategy at this point. Only one feature of a hypothesis is added at a time and tested.

Much information is obtained by conservative focusing and redundancies are avoided. The risk of securing no information is under control; in fact, one always gets some useful information. The only real disadvantage is that information which is not arranged in orderly fashion might involve a long search for the needed data. Thus, part of the success of a subject using conservative focusing depends upon how well organized his knowledge is.

d) Focus gambling. The initial steps taken by a subject who uses focus gambling as a strategy are the same as those of one who uses conservative focusing. But, he is not satisfied with adding one part of a hypothesis at a time and testing; he adds two or more parts at a time and gambles. The strategy is not quite an all-or-nothing one, but it comes very close in case the omnibus hypothesis fails. The only thing to do then is for the subject to shift strategies and fall back on successive scanning.

However, the gamble may pay off. The omnibus hypothesis may lead to attainment of the objective. There is much gain then, of course, for the number of trials has been reduced to a minimum and much time has been saved.

THE STRATEGIES IN USE

IMPORTANCE OF INITIAL FOCUS

Several years ago, the writer did a pilot study in which about ninety secondary-school teachers were involved. Two tasks, or problems, were presented to the subjects, and they were asked to find a solution. The objective at the time was to secure information on these teachers' concept of a mathematical solution, but other interpretations of the evidence turned out to be more interesting. This is true, particularly in view of the foregoing discussion.

Situation 1. One problem was the following:

> From a two-digit number, N, subtract the number whose digits are those of N interchanged. If the difference is a positive cube, how many permissible values of N are there?

However innocent this problem appears, it is not a run-of-the-mill problem. The method of approach, that is, the initial focus, is critical. Practically all of the teachers classified the problem as a "digit problem." All right; there are standard methods for solving a problem so categorized. These were applied; let t and u represent the tens' and units' digits, respectively; then

$$N = 10t + u$$

$$\text{a positive perfect cube} = (10t + u) - (10u + t).$$

Quickly perceived is that the tried-and-true methods fail. There is no place to go from this point, for at least two more equations are needed for them to be successful. A small number of papers were discarded because of errors in finding a simpler form for $(10t + u) - (10u + t)$. It is doubted that this was failure of any strategy. Rather, such errors seemed to be the result of haste or carelessness. The remaining papers fell nicely into three groups.

There were those whose initial focus was all right. These papers showed that the persons got as far as

$N = 10t + u$
$x^3 = 9(t - u)$
$x > 0.$

But, unfortunately, once they had made a focus, they were not flexible enough to change it when it led nowhere. The papers of these subjects were covered with equations and tag ends of scribblings. They formed only one hypothesis – secure enough equations to solve for all of the unknowns. One might categorize their strategy as successive scanning, where there was just one hypothesis to scan. This was done over and over by members of this group. There was no change of focus with respect to the problem. Failure did not discourage them from trying the same hypothesis again and again.

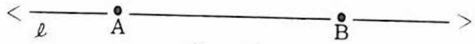
The second group's papers showed fewer scribblings. There was hardly any writing on these papers after the initial symbolizing of the problem situation. Inferred from what evidence was available was that these subjects realized that the initial categorization of the problem as a digit problem was not quite correct. They seemed to be flexible enough in their focusing on the problem to change their focus. It seemed, however, that they were at a loss to formulate another hypothesis or another strategy. A possible interpretation is that these people had a limited concept of a mathematical solution. Geometric proofs and the results obtained by solving equations qualified all right, but the result found by trial and error or by a verbal argument did not.

Members of the third group not only realized that a shift in initial focus was necessary but they also had a better conception of what constitutes a mathematical solution. Some adopted a strategy of successive scanning, trying different pairs of numbers in succession. Others clearly used conservative focusing, for they not only tried pairs of numbers but also took into account the positive perfect cube datum. Those with the most elegant solutions used a highly sophisticated conservative focusing strategy. A few used focus gambling, but more due to the pressure of time rather than to preference.

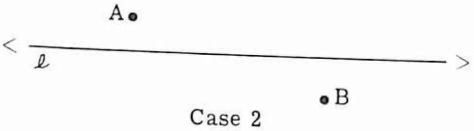
Situation 2. The second task presented to these teachers was a straight-forward geometry problem:

> Find, on a straight line, a point such that the sum of its distances from two given points is the least possible.

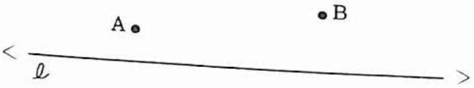
The first step in any geometry problem is to draw a figure and introduce some notation. If more subjects had followed this advice, there would have been more correct solutions to this problem. As it was, only one-sixth of the subjects were able to arrive at a solution. Just as was the case with the algebra problem, here, too, the difficulty was inflexibility with respect to the initial focus.



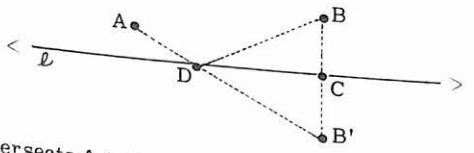
Case 1



Case 2



Case 3



Let us take the key idea of the problem, the sum of the distance from two given points is the least possible, and use it as an initial focus. Now, as one proceeds to make a figure and introduce notation, he must give up this initial focus for a moment and shift to focusing on the two given points. The problem says nothing about placing them; a decision is required. Three possibilities, or hypotheses, appear: Case 1, both points are on line ℓ ; Case 2, the points are on opposite sides of ℓ ; and Case 3, the points are on the same side of ℓ . A shift of focus back to the least possible distance-focus immediately takes care of Case 1 and Case 2, for the shortest distance between two points is the length of the segment joining them. Now, keeping these results in mind, turn to Case 3 and ask "How can I use this idea (the length of the segment)?" A few moments' thought suggests reflecting one of the points in the line ℓ , yielding point B'. Segment $\overline{AB'}$ intersects ℓ in D, and D is the desired point. For triangle BDC is congruent to triangle B'DC, and consequently BD = B'D. Therefore, BD + DA is the least possible distance.

These two problems have emphasized the importance of initial focus. More than that, discussion of attempts at solutions made clear how necessary it is for a problem solver to be flexible enough to shift his focus as the process of solving the problem develops. In each problem, concepts related to that involved in the problems themselves were very important. It seems that, at least in mathematics, knowledge of these subsidiary concepts is crucial to success. The strategies used seemed first to be that of simultaneous scanning with a transfer to conservative focusing after a promising "lead" had been discovered.

SIMULTANEOUS OR SUCCESSIVE SCANNING?

The next problem situations are relatively independent of initial focus, depending much more upon a choice of strategy. The mathematics curriculum puts considerable emphasis today on a search for patterns. The following situations are of that sort.

Situation 3. Divide your paper into three columns. In the first column there is a string of letters, and in the second column there is another string of letters. Now, we shall use these two strings of letters to make a new set of letters in the third column. How were the letters in the third column obtained? Now, try to write the set of letters that goes into the third column for these two sets. What goes into the third column for the following two sets?

K, L, M, T, W, Z	L, M, T, X, Y	

K, L, M, T, W, Z	L, M, T, X, Y	L, M, T

A, C, E, P, Q	A, C, L, P, Q, R, S	

T, U, V, W	T, U, V, W	

Here is another example. What would you write in the third column?

S, A, M, E	N, O, T	

Is it true that, given sets in the first and second columns, there is always exactly one way to do the third column?

What are the results if we interchange the sets in the two columns? Fill in the third column and compare the results with what you got before.

L, M, T, X, Y	K, L, M, T, W, Z	
A, C, L, P, Q, R, S	A, C, E, P, Q	
T, U, V, W	T, U, V, W	
N, O, T	S, A, M, E	

What did you find?

Situation 4. Now let us form a third column set in a different way. As before, we start with sets in the first and second columns. How was the

A, C, E, G, H, J	B, C, D, E, F, G, T	A, H, J

third column set obtained? Now, try to write the third column set, starting

F, A, M, I, L, Y	E, A, R, L, Y	

with these two sets. Let us try two more special situations. Is it always

A, B, C,	D, E, F, G	
A, B, C, D	A, B, C, D	

possible to obtain exactly one set, given sets in the first and second columns?

Let us see what the results are if we interchange the sets in the two columns. Compare the answers with those you had before. What can you say?

B, C, D, E, F, G, T	A, C, E, G, H, J	
E, A, R, L, Y	F, A, M, I, L, Y	
D, E, F, G	A, B, C	
A, B, C, D	A, B, C, D	

We have worked with strings of letters in two different ways. That is, we used two different patterns to obtain the set in the third column. Compare these two patterns. Do they have the same properties? How are they different?

DISCUSSION OF SITUATIONS 3 AND 4

The last two problem situations are quite different from the first two. Except that they were not quite straight-forward, the first two problem situations are representative of the mathematics program before World War II. The last two are representative of types of problems that arise in the new mathematics program. The algebra and geometry problems have for their objective getting an answer. After the answer has been obtained, there is little one can do with it, except perhaps to refine the method of solution. It is not so much that a concept is in the process of being attained as it is that pre-established concepts serve as mediators in solving each problem. Consequently such problems would be reasonable to present only after there was some evidence that subjects had been at least exposed to the necessary mediating concepts.

On the other hand, the third and fourth problem situations are representative of situations intended for seven- to nine-year-old subjects. The intent is that through such situations a foundation is laid for attaining rather deep mathematical concepts – those of closure and commutativity in this case. A battery of situations of this sort can be devised that has a bearing on ordinary arithmetic. Of course, letters would not have to be used. One could use the boys and girls in a class, or colored counters, or books, or any other sort of material available. An effort would be made to use materials appropriate for the age level of the class.

It is not clear to the writer that mediating concepts are brought into play. Rather, it seems that solving these problem situations requires use of a strategy, one that is effective in attaining concepts. Pattern-problems such as these seem to yield to a successive scanning or a simultaneous scanning strategy. Or, one uses a combination of the two strategies; simultaneous scanning for gross eliminations of parts of the strings of letters; and successive scanning for determination of fine discriminations.

TWO ITEMS OF RESEARCH

Important to the mathematics program is knowledge concerning what types of data in problem situations present difficulties. Such knowledge influences preparation of classroom materials in two ways. Learning experiences can be provided that are intended to facilitate attainment of a concept, and information gleaned from research studies can be used to schedule concepts according to age levels.

An attempt by one person to report on all of the research studies concerned with problem solving would be foolhardy. A bibliography is very long. Two doctoral studies have been singled out for reporting. One reason is to underscore the fact that much relevant research gets lost or overlooked through non-publication in journals.

POST STUDY

In 1958, Post did an experimental study of six factors involved in understanding problems. His subjects were fifth- and sixth-grade pupils in schools in the Metropolitan New York section of the country. The definition of a problem in the study was "a quantitative situation described in words . . . but in which the arithmetical operations that lead to the answer to the question are not explicitly indicated" (Post, 1958, p. 15). The six factors were: 1) size of numbers; 2) superfluous numerical data; 3) familiarity of setting; 4) number of steps; 5) type of operation; and 6) symbolic terms.

Factor 4, number of steps, requires some words of explanation. Post defined a two-step problem and a three-step problem as follows:

> A step in a problem solution . . . was a simple performance of one of the four operations of addition, subtraction, multiplication, and division. Therefore, a problem is (a two-step problem) if and only if the solution of the problem requires a single performance of two different operations or two performances of the same operation (Post, 1958, p. 30).

A three-step problem is defined in a similar way.

Factor 6, symbolic terms, refers to the question of whether using a name for a number such as "twenty-six" in place of "26" interferes with a subject's ability to solve a problem. The names of the other factors are self-explanatory.

Two forms of a test were prepared by Post; these forms were judged equivalent by a jury of experts and then tested further in a pilot study. Familiarity of setting and non-familiarity of setting was obtained from an inventory of such settings suggested by elementary-school teachers. This information was supplemented with evidence obtained from the pilot study pupils.

The study design was such that an analysis of variance technique could be used. The error term "was composed of the unconfounded four- and five-factor interactions as well as the six-factor interaction" (Post, 1958, p. 85). The two forms of the test were treated as two replications of the same experiment, for examination of the test results from the two forms led to doubt that the data from them could be combined. The schools represented two different socio-economic neighborhoods; the subjects were chosen from non-homogeneous classes; and small sample theory analysis was used. Correctness of solution was determined from examination of the work written by a pupil rather than from the correctness of the answer. For example, a pupil's solution might be categorized as correct even if his answer was incorrect. Another way of saying this is that errors in arithmetic computation were disregarded.

A hypothesis was rejected if the results from both forms of the test indicated rejection; it was accepted, similarly, if both forms indicated acceptance. No conclusion could be stated if one form indicated rejection and the other form indicated acceptance.

The hypotheses on the factors of size of numbers and number of steps fell in the no conclusion category. So far as the latter (number of steps) is concerned, it might have been that there was a strong interaction between number of steps and type of operation, for the hypothesis concerning the latter was rejected at the one tenth of one percent level of significance. That is, problems involving multiplication and division were found to be more difficult for the sample subjects to understand than problems involving addition and subtraction.

Hypotheses rejected were those concerning superfluous numerical data, familiarity of setting, and the already mentioned type of operation. The hypothesis concerning symbolic terms was accepted. In short, expressing numerical data in words or in digits had no effect on pupil's understanding of such problems.

BECHTOLD STUDY

The 1965 study of Bechtold consisted of a comparison of an experimental group with a control group in which the former received instruction in work with problems containing superfluous data and the latter did not. Two units from the ninth-year program, inequalities and applications of equations to falling body problems, were chosen for instruction and testing. The subjects were 14- to 15-year-old students in New York City and the surrounding Metropolitan region schools.

The hypothesis that instruction in handling problems containing superfluous data (in both inequality and falling body problems) had no effect on developing problem-solving ability was rejected at the one percent level of significance. The results with respect to variables of sex, intelligence, and age indicated no significant differences with respect to sex and intelligence. The data indicated, however, that younger students seemed to benefit more from instruction than older students.

RELATION OF THE POST AND BECHTOLD STUDIES TO CONCEPT ATTAINMENT

It is clear that if there is such a thing as a concept of problem solving, then it is a very complex concept. Moreover, success with problem solving is dependent upon mediating concepts that in themselves are complex. The studies of Post and Bechtold, for example, underscore the importance of instruction, of exposure of pupils to many different sorts of problems or learning experiences in order that they have opportunities to attain necessary subsidiary concepts.

The fact that familiarity of setting and superfluous data were significant factors in Post's study seems to indicate that pupils may have had no previous encounters with such problems. Hence, they would have had no opportunity to form a concept of such problems, that is, to categorize them and develop strategies to use in their solution. Since multiplication and division in problems turned out to be significantly more difficult to understand than addition and subtraction, one infers that pupils do not conceptualize the former as well as the latter. It may be that the methods of teaching used with multiplication and division failed to develop firm concepts. As a result, pupils were unsure in their categorizations when these operations were present in problems.

Bechtold established that instruction in how to handle problems containing superfluous data led to greater success with such problems than that of a control group – at least for ninth-grade algebra students. A teacher is prone to extrapolate from this evidence and conclude that such instruction, appropriately designed, is good for almost all age levels. Although the study does not represent conclusive evidence, still it is encouraging in that statistically significant differences resulted. So often an experiment that compares two procedures in instruction ends with one group doing about as well as the other on tests; there are no statistically significant differences in the two sets of results. The unexpected significant differences in Bechtold's study leads one to go beyond the statistics and say that experience with the analysis of problems with superfluous data significantly affects outcomes.

From the definition of a concept given at the beginning of this paper, one sees that a person does not form a concept just through maturation. There must be experiences in his environment that make his attempts to understand the environment too complex unless he forms classes – unless he conceptualizes. This is particularly true of mathematics, one of man's sophisticated inventions. Perhaps a child develops "naturally" some concept of counting, even inventing his own names in some instances. But as there are encountered larger and larger groups, the effort to remember number names becomes too much, and, if historical evidence is valid, there is a turning toward forming classes, groupings, and a numeral system is born. From this point on, present day culture dictates that intensive exposure to a variety of experiences must leap-frog the historical pattern of development in mathematics. These learning situations cannot be left to chance; they must be provided; and they must be planned carefully. Experimental evidence, both in the field of psychology – learning theory – and mathematical education, furnishes guidelines for such planning. In turn, experimental evidence affirms or denies that the planning is effective. The Post and Bechtold studies are examples of this two-pronged approach to a mathematics curriculum.

SUMMARY

So it is that human beings tend to deal with classes of things instead of individuals in order to make some sense out of their environment. By forming such classes cognitive strain is reduced as well as the burden on memory. To form these classes, or categories, or sets, a person looks for cues, or, if you like, for characteristics that serve to distinguish things eligible for membership in the set from those that are not eligible. The point is that these categorizations are inventions, and this is particularly true in a complex body of knowledge like mathematics.

To solve a problem in mathematics, subsidiary concepts serve as mediators. Although initial focus and type of strategy have important roles, an individual will have little success in problem solving unless he has a firm grasp of the mediating concepts. Mathematical education speaks in terms of a student mastering a concept so that it is a part of himself; a psychologist would speak in terms of a student internalizing a concept. No matter what language one uses, it is certain that some of the failures in mathematics instruction are due to an instructor assuming his students have understood a mathematical concept at a high level of operational thinking when in reality they have a much lower level of mastery. They cannot apply the concept in an unfamiliar setting. As experimental evidence accumulates from investigations of how people learn and from studies in a curriculum context, mathematical educators are on the one hand becoming more cautious and on the other more daring. Applying the evidence, they have changed the content of the curriculum and reduced the amount of pure drudgery that used to be the lot of every mathematics student. Assessing recent evidence, they have some doubts whether all students master the early introduction of concepts.

Already said is that people categorize the things in their environment. A categorization might result in an initial focus on a problem. Then certain hypotheses are formed that are related to this initial focus. Decisions on these hypotheses serve to validate the hypotheses.

> The sequence of decisions made by the person en route to attaining the concept [which may be a problem solution or a pattern] . . . may be regarded as a strategy embodying certain objectives. These objectives may be various in kind but in general one may distinguish three kinds of objectives: a. to maximize the information gained from each decision and test of an instance; b. to keep the cognitive strain involved in the task within manageable or appropriate limits and certainly within the limits imposed by one's cognitive capacity; and c. to regulate the risk of failing to attain the concept within a specifiable time or energy limit. . . (Bruner, et al., 1956, p. 234).

In the examples presented of problem solving or of attaining a concept in mathematics, the idea of initial focus and flexibility with respect to one's focus is of great importance. After isolating a key idea in a problem, a student has a choice of strategies – simultaneous scanning, successive scanning, conservative focusing, or focus gambling. A skilled problem-solver probably uses a combination of simultaneous scanning and conservative focusing. A beginner is likely to use successive scanning. If time harasses a problem-solver, there is nothing to do but turn to focus gambling as a strategy. The situation and the individual (his own experiences, battery of concepts, and psychological make-up) determine which strategy will be used.

The investigations of Dienes (1963) at Harvard as well as that of Dienes and Jeeves (1965) in Australia lean heavily on the work of Bruner and Piaget. However, their work goes beyond that of their predecessors, making use of new research results and suggesting refinements of the theory of learning upon which Bruner based his work. An idea that Dienes and Jeeves, separately and together, experimented with is the development through "mathematical games" of the concept of isomorphism of structures. Involved here is the recognition of patterns and noticing the similarity between two sets of data and the structure that can be applied to both.

Suppes (1965) has done a great deal of work with the major emphasis on learning models for relatively non-complex mathematical concepts. He characterizes his work by saying that it is based upon a stimulus-sampling theory. An interesting assertion is that the hypothesis-strategy language of Bruner and his associates can be related to stimulus-sampling terms: ". . . a strategy . . . corresponds precisely to a state of conditioning and a hypothesis to the conditioned stimulus sampled on a given trial . . ." (Suppes, 1965, p. 65).

Some psychologists doubt the value of a mathematical psychology approach to models for human learning, but there is no disagreement that much work needs to be done with complex concepts in order to learn more about individually different approaches to attaining mathematical concepts. Meanwhile, the mathematical educator makes use of available knowledge, applies it as best he can to the mathematics program, and gathers experimental evidence to see the effect. It is impossible to wait until definitive psychological research on learning has been done, for the curriculum exists and must be modified to fit twentieth century activities. One implication is clear: new psychological evidence should be brought to bear on the curriculum as quickly as possible. It is in the latter direction that mathematical education is turning.

NOTE: Situations 3 and 4 are modeled after similar materials that appeared in the Nuffield Foundation Mathematics Teaching Project Bulletin No. 1, November 1964, pp. 4-7.

REFERENCES

Bechtold, C. The use of extraneous material in developing problem-solving ability of pupils in Algebra I. Unpublished doctoral dissertation, Teachers College, Columbia University, 1965.

Bruner, J. S., Goodnow, Jacqueline J., and Austin, G. A. A study of thinking. New York: Wiley, 1956.

Dienes, Z. P. An experimental study of mathematics-learning. London: Hutchinson, 1963.

Dienes, Z. P., and Jeeves, M. A. Thinking in structures. Psychological Monographs on Cognitive Processes. Vol. 1. London: Hutchinson Educational, 1965.

Klausmeier, H. J., Davis, J. K., Ramsay, J. G., Fredrick, W. C., and Davies, Mary H. Concept learning and problem solving: A bibliography, 1950-1964. Technical Report No. 1, Research and Development Center for Learning and Re-Education, C-03, OE5-10-154. Madison: University of Wisconsin, 1965.

Post, R. A study of certain factors affecting the understanding of verbal problems in arithmetic. Unpublished doctoral dissertation. Teachers College, Columbia University, 1958.

Suppes, P. On the behavioral foundations of mathematical concepts. In L. N. Morrisett and J. Vinsonhaler (Editors), Mathematical learning. Monographs of the Society for Research in Child Development, 1965, 30, (1, Whole No. 99), 60-96.

PART III

Instructional Materials and Media

THE MEDIUM of instruction can be considered as the interface between the learner and the subject matter being learned. It necessarily affects the rate of learning and imposes limitations on the scope and sequence of the instructional subject matter. Instruction by television has been criticized as impersonal and encouraging passivity. Yet, according to Wilbur Schramm in "Instructional Television Around the World" instructional television can, when properly used, provide for active learning experiences. In a series of case studies, Schramm reports on the use of television in widely varying sections of the world—from the affluent suburb to the developing nation. Properties and attributes of the media are discussed with reference to the school system, the teacher, and the student. Television is examined as a potent mechanism of change. Schramm concludes that this medium works best when it is integrated into an instructional system which includes specialized materials, activities, and teaching techniques. But the key to the system is a local supervisor through whom students can interact with one another.

Research dealing with the process of instruction is examined by Arthur Lumsdaine who points out that problems associated with instruction are in fact inseparable from problems dealing with learning. The close relationship between research on instructional media and the more general field of basic psychology of learning must be understood by persons purporting to do research on instructional media. Among the general areas identified for research on the effects of instruction is the "far underdeveloped research field of constructive and imaginative attempts to improve on the basis for determining what ought to be taught." Lumsdaine argues that determining the effects of present instructional media should precede the "more elegant question of what specific variables account for these effects." He takes the position that effective instructional media are a prerequisite to meaningful basic research on these media. This implies programs of development before research and suggests a basis for priorities in research and development. "Only when something effective has been developed, do we then have an adequate tool with which to conduct useful research."

The question of cost efficiency which is usually not considered by educators is explored. Such efficiency frequently implies a measure of the level of proficiency achieved for a given effort with, as Lumsdaine suggests an approach in which the time required to achieve a given minimum level of proficiency is used as the dependent variable. Such a proposal has significant implications for the design of instructional objectives and instructional research as applied to selected subject areas.

Lumsdaine concludes with a discussion of the standards or lack of standards used in the design and reporting of research studies dealing with instructional media.

Matching the instructional strategy to the capabilities of a learner requires a tutorial model of instruction. Lawrence Stolurow proposes an instructional system in which methods and materials are individualized for the learner on the basis of the continuous matching of methods and materials with the aptitude, personality, and knowledge of each learner through the use of feedback contingencies both prior to and during instruction. Such individualization can be achieved only in Computer Assisted Instruction. Stolurow summarizes a general idiographic contingency model of a cybernetic, computer based instructional system called SOCRATES. "It [this model] is based on the assumption that individualized instruction should take a prescriptive-corrective approach to the use of psychological information in making decisions about teaching in a tutorial context." The logic of instruction is determined according to certain rules whose application is contingent on variables monitored throughout the instructional sequence. Results based on the idiographic model are presented with reference to the kinds of decisions made in individualizing instruction via the computer.

Instructional Television Around the World

Wilbur Schramm
Stanford University
Stanford, California

WE HAVE now in hand the largest body of comparable and detailed data yet assembled on instructional television throughout the world: twenty-three case studies of the use of new educational media in eighteen countries, most of them including television.[1] I will review three of these cases in some slight detail, cite a few others, and then try to sum up what I think the studies say about the requirements for using instructional television effectively.

One point I shall not belabor is that instructional television, used well, clearly does work. It works over a wide spectrum of tasks and a variety of situations. It works in developing countries and countries that are highly industrialized. It works sometimes very well and sometimes ordinarily. To illustrate what these cases say about how it works and under what conditions it appears to work well, I will enumerate three instances.

CASE 1: SAMOA

The story of instructional television in Samoa really begins with the appointment of a tough, determined, little governor, H. Rex Lee, in 1961, but the whole story is 60 years older than that, and deals mostly with American neglect of its wards on the Samoa islands. The school system was only one symptom of that neglect, but it was a vivid one.

The United States had committed itself to offering Samoans an education equal in every way to that on the mainland. Yet the Samoan schools were in one-room, open fale huts, where two classes competed with each other, fortissimo. The instruction itself was the most traditional rote learning. I have been in these village schools and have heard the kind of choral chanting that one becomes accustomed to in primitive schools throughout the world. The students repeat back – chant back – what the teacher gives them; indeed, this is the only kind of teaching possible when the teacher is untrained, little educated, and not at all confident of his grasp of the subject matter. He doesn't dare encourage questions or let the students wander away on their own intellectual excursions. The teacher says, "Class, what am I doing?"

Like one, the class chants back, "You - are - standing - on - the - mat."

"Class, what do I have in my hand?" the teacher questions.

"You - have - a - piece - of - chalk," the class replies.

Once I rose, smiled as disarmingly as possible at the teacher, and said "Class what am I doing?"

"You - are - standing - on - the - mat," they responded promptly, although I was far away from the mat.

I picked up a book. This puzzled them; they could see it wasn't chalk. "What do I have in my hand?"

One boy ventured: "You - have - a - piece - of - chalk."

They were quite unable to generalize much upon the responses they had learned in their educational catechism.

The results of this kind of thing were predictable. The schools were supposed to be the equals of mainland schools, but children graduated from the twelfth grade testing on the average from fifth to ninth grade in different subjects. The teaching was supposed to be in English, but little of it was, and many of the teachers could not speak the language understandably. The content was derivative from mainland texts, and so the children learned their reading vocabularies from accounts of commuter trains, milk men, and snowball fights. I saw one elementary-school class being drilled to remember that "The 'Mona Lisa' is in Milan," although the significance of that knowledge to a Samoan child was not immediately obvious. And over everything hung the pall of traditional rote instruction.

Now, how do you transform a traditional school system like that, in a few years, into a modern problem-solving, inquiring school system? How do you live up to the promise and plan to teach in English, with so few English models? That is the problem Governor Lee faced. I am begging the question of whether the level should have been raised in a few years, rather than a century; whether it was important to have English well taught and well used; whether an excellent 12-grade school system was what Samoa most needed. But given that policy goal, how would you proceed?

What were Governor Lee's alternatives? He could replace all three hundred Samoan teachers by qualified teachers from the mainland, if he could recruit them. But this would have been a politically, psychologically, and ethically impossible solution. He could bring in perhaps one hundred teachers from the mainland. This would displace less than one-third of the native teachers and stir up less trouble, but would run head on against one of the oldest Samoan norms – that all go forward together. He could institute a greatly increased teacher program. But this would require sending prospective Samoan teachers to the mainland for some years – lest they merely perpetuate the quality of the old school system – and its effect would not be felt for perhaps 15 years. Or he could use television to share a small amount of expert teaching over many classrooms. And this is what he decided to do.

But television itself was not enough. The people who studied the problem saw very soon that what they had to create was really a new teaching-learning system. They had first to consolidate the schools and provide adequate learning conditions. Then they had to create a sort of teaching team, in which an expert teacher would give the core of the curriculum by television and a native classroom teacher would be taught to build a context of learning activities around the broadcast; and still another teacher would prepare study materials and exercises to fit into the plan. Make a mental note of this, because the same conclusion is being reached all over the world – that whereas pupils learn a great deal from television alone, they learn a great deal more when it is built into such a teaching-learning system as I have described. Nowhere in the world is television being used entirely alone and effectively for any major instructional task. Always it is integrated into some kind of system in which the studio teacher, the classroom teacher, and sometimes others cooperate in stimulating the maximum amount of learning activity in the students.

Governor Lee persuaded Congress that our neglect of Samoan education should be repaired. A remarkable installation was built at Pago Pago, one of the finest instructional television installations in the world, with six open-circuit channels, ten videotape recorders, transmitters on a mountain top, and television teachers, engineers, and producers recruited from the States. Classes began to move into the new school buildings in 1964, and television began to carry the burden of teaching beginning in the autumn of that year.

How is it working? What evidence is there that television can raise a primitive to a modern system – Beeby's (1966) Class I and II schools to his Class III and IV schools – in a short time?

The evidence isn't all in, yet, but the signs are encouraging. The Samoan elementary schools are working very well, indeed. The children are speaking much more and much better English. The television classes are live and active; the pupils talk to the teacher and answer her questions just as though she were in the room. There is much less rote practice. Children are asking questions. Some schools are making barometers and undertaking other special projects. I saw one sixth-grade teacher bring in a lizard for the class to study. They talked about the little beast, what he ate, who ate him, and worked around to the balance of nature. This could not possibly have happened only a few years ago. I can't say that the high schools – which are using, mistakenly or not, very large classes – are working as well as the elementary. But on the whole, it is a very encouraging and hopeful experiment, one that is worth watching.

Pause briefly, before we review the next case, to reflect on what seem to be the basic elements of the Samoan case: a problem, large, urgent, important enough so that many groups can rally to solve it; adequate support, administrative and financial; and the introduction of television, not as a separate element, but as a part of a teaching-learning system integrating all the teachers and all the classroom activity.

CASE 2: HAGERSTOWN

Let us turn now from the fale huts and palm groves of Samoa to the green fields and active industries of Maryland. The United States experiment most often linked to Samoa is Hagerstown. Most of you are familiar with this system, and I shall not enumerate what you more than likely already know about it.

As you probably know, there were three chief reasons for introducing television into the Washington County school system, centering on Hagerstown. One was that school officials were anxious to make expert teaching of art, music, and science available throughout the elementary schools, and to offer certain advanced courses, such as college-level mathematics in high school. To do this with resident teachers would have added a great deal to the budget, and would have presented severe recruiting problems. Therefore, they wondered whether it was not possible to share expert teaching in these subjects by means of television, at a lower cost, or at least no higher cost, than by hiring the necessary specialized teachers. In the second place, they wondered whether in the coming years of expanding enrollments there would not be some advantage in designing schools, not with many rooms about the same size, but rather with a central large room or rooms for viewing televised classes, and a number of small rooms around the perimeter of the building. And, finally, financial help was available – $1 milion from the Ford Foundation and over $300,000 in equipment from manufacturers.

This was in 1956. Hagerstown decided, as did Samoa later, that it needed the equivalent of six channels to teach the core of an entire curriculum. Because open-circuit channels were not available in that number at any one point in the United States, it was decided to go closed-circuit. Coaxial cables were leased from the telephone company, and extended gradually over the next few years to every school in Washington County. For more than a half dozen years, now, every student in the system has received the core of his curriculum from television.

In Hagerstown, as in Samoa, a studio teacher and a classroom teacher operate as a team to teach content previously decided upon. The studio teachers are chosen from the regular teaching corps of Hagerstown, and are changed frequently. Because all the teachers in the system are qualified and well trained, there has been less difficulty, in the Hagerstown than in the Samoa experiment, in getting the classroom teacher to play his new role expertly. But two elements are generally credited with the smooth transition in Hagerstown from the older plan to the television teaching team. One of these is the firm support given the project by Superintendent William Brish. In a sense he played the role in Hagerstown that Governor Lee played in Samoa. The second is the insistence throughout on involving as many teachers as possible and thus reducing the potential threat of the new method. Only about 5 percent of all the Hagerstown television teaching is taped, and very little of it is brought in from outside. Thus teaching on television is treated as an ordinary function of team teaching, not specially rewarded, and passed around among promising candidates.

A considerable amount of research has been done at Hagerstown, and some of the results of introducing television have been spectacular. Remember, now, that this is not a deprived school system, nor was it in 1956. It is a rather average American system. But in Hagerstown junior high schools during the first four years of television the average performance of students on a standardized test of arithmetical concepts rose from the thirty-first to the eighty-fourth percentile, measured against national norms, and on a standardized test of problem solving it rose from the thirty-third to the sixty-eighth percentile. In other words, before television, two-thirds of the junior-high-school pupils of the country scored above Hagerstown children on these arithmetic tests; after television, two-thirds scored below Hagerstown. In the fifth grade, during the first year of television teaching, the average Hagerstown pupil gained 1.9 years, rather than the one year represented by national norms. In the year before television was introduced, every one of the grades three through six averaged below national norms in arithmetic. Within 2 years, each grade averaged above the national norms. We have spoken only of arithmetic. Not all results were as good as these, but the gains were by no means limited to arithmetic. Perhaps the most significant verdict was turned in by the school board of Washington County who, in 1960 when the Ford grant came to an end, studied the results, gave all teachers and administrators a chance to express their opinions anonymously, and then voted without a dissenting voice to take over the approximately $600,000 annual cost of the television in their own budget.

CASE 3: NIGER

Both Hagerstown and Samoa, one might say, profited by American know-how in television. Suppose now we turn to a country that has had little or no help from the United States in television. Niger has only sixty-six fully qualified teachers – meaning teachers who are secondary school graduates and have had teacher training. There are some hundreds more who have had the equivalent of junior high school education, and still more who have had only 5 or 6 years in school. But there is still no hope for adding to the resource in highly educated and trained teachers until the country's higher priority needs for educated people have been met. This creates an extremely difficult situation because Niger, which has had only between 5 and 10 percent of its school-age children in the classroom, is now trying to double and soon to triple its school population, and especially to strengthen its secondary schools. The problem was, then, how to find the necessary number of skilled teachers.

Acting with the advice and help of the French government, Niger developed a plan for using expert teachers in the studio and monitors in the classroom. Unlike Samoa and Hagerstown, which began with all the elementary and secondary grades at the same time, Niger began with the first grade. It began very cautiously, in fact, with only two schools, and with closed-circuit television. By the time the country's new open-circuit television station went on the air, the first-grade experiment had shaken down, and the television was expanded to the second grade. The plan is to add one grade a year.

Like Samoa and Hagerstown, Niger has found it works best to create in effect a teaching team in the studio and the classroom, and like Samoa it has found that the weakest link in the team is the classroom teacher when that individual is not adequately trained. Using elementary-school-trained monitors in the classroom, the supervisors of the Niger project have tried short in-service training courses, and have decided more intensive special training is needed. One of the ways Samoa tried to solve the problem of helping the relatively untrained classroom teacher to learn his new role was by putting into every school an American principal, who worked with his teachers and served really as a supervisor. They also brought all the teachers to a six-week summer workshop each year, and conducted daily in-service training by television. No easy way has been found anywhere to solve this particular problem, but it has reappeared again and again in developing countries where instructional television has been introduced.

One of the distinctive features of the Niger experiment has been the continuing presence of a three-man research team supplied by France – a psychologist, a sociologist, and an ethnologist. These people are studying very closely the effect of television teaching on African children who have never before seen television, and, most of them, have never before spoken French, which is the language of instruction. Their psychological and cultural reports are not yet available, but they have measured the learning results of the first year. As might be expected, the results were better in some subjects than others, and somewhat better for children who were a year older than the others. But as a whole, the television classes averaged well above the average performance of non-television classes. We must say of Niger, as of Samoa, that all the evidence has not come in, but the outlook is highly encouraging.

SOME NOTES ON OTHER CASES

We will deal with several additional cases in a more brief manner.

You have noticed that in all three of the cases I have mentioned, generous financial help was available, and in the case of the two developing countries expert counsel was also freely offered. What would happen now if a developing country tried to go it on its own?

It isn't a recommended procedure, but it has been done. It was done, for example, in Arequipa, Peru. In that high Andean city, a group of school teachers, members of a religious discussion group, decided to undertake a community project. Looking around them, they saw that hundreds of children were unable to gain entrance into the schools because of shortage of places for them. Hundreds of others had left school after 3 or 4 years. They decided to begin with a group of the school-leavers who had gone to work as domestic servants. Suppose that the employers of these children would let them watch television an hour a day, they said. And suppose that the proprietor of a small local television station would give some free air time. And suppose that we were to do the teaching, and could get help in preparing learning materials. Well, all these things came true. Although with some objection, the employers let their domestic servants watch television; the station offered up to three hours free time; the teachers donated their time; the priest who was head of the discussion group supervised the curriculum. And a television school for school-leavers went on the air without a single paid employee. Since that time, the project has immensely expanded, now has a government contract and some paid employees, and offers a varied group of out-of-school courses, adult courses, and so forth. It has been almost fiercely local, resisting rather than inviting outside advice and help, and this has not been wholly to its advantage. But at least the Arequipa story shows that the do-it-yourself spirit is not entirely missing from instructional television in the new countries.

We have reviewed several uses of instructional television in school. The Arequipa project was mostly for out-of-school use. There have been a number of other projects of that kind, aimed at extending educational opportunities beyond the school – for instance, Telescuola which provides secondary education for Italian children who do not have secondary schools in their areas; and the Chicago Junior College which for nearly 10 years now has offered a complete junior college curriculum by television, with the home television students doing consistently better than classroom students on the same examinations. Television has been used to help teach literacy – in places like the Ivory Coast where it has helped to train native workers to read, write, and speak French, and do simple arithmetic so that they could be made middle-level supervisors; and in Italy where about 15,000 adults annually register to study reading, writing, and arithmetic, with the aid of the famous program series, ''Never Too Late.'' The trend of the evidence in literacy experiments, though, is that, whereas television can assist greatly by providing visual material, by motivating a group, and bringing a certain amount of expertness, the real key to what happens lies in the ability of the group supervisor. A straw in the wind is that Italy, which at first used only monitors in its viewing groups, now uses certificated teachers – usually working outside school hours – for that purpose.

One more note: from almost everywhere instructional television has been used comes word that it is an excellent vehicle for in-service teacher training. The mere example of excellent teaching televised into a teacher's own subject is itself effective. As some observers in Samoa remarked, this may have been the first time that some of the native teachers had ever seen a well-made lesson in their own grade level and field. Furthermore, to prepare the teachers for teaching an unfamiliar subject or method, like the new math, television has proved an excellent in-service training tool. It was used for that purpose in Colombia, for example, and the results were carefully measured. The amount of learning was very high, but with teachers as with pupils the total amount depends greatly on what happens on the receiving end. That is to say, whereas a considerable amount was learned from television alone, a great deal more was learned if the teachers met in groups and discussed the broadcast lesson, and still more if the groups had competent supervisors to direct the discussion.

We have said enough now to indicate that instructional televison, used well, can contribute toward the solution of a great variety of educational problems – the need to raise instructional level quickly without a teaching corps adequate for the purpose, making up for the scarcity of specialized or highly trained teachers, extending the school beyond its physical boundaries, helping to teach adult literacy, and providing in-service training for teachers. This is not, by any means, to say that it always works well. I could cite some very disappointing stories to balance the encouraging ones I have mentioned. But I believe we know enough now to advance a set of propositions as to what are the conditions under which it is likely to work effectively. Let me suggest some of these.

CONDITIONS OF SUCCESS

(1) A problem.

After twenty-three case studies, I devoutly hope that the next time someone comes to me for advice on instructional television, he comes with a problem and not with a piece of technology. This is why some of the more spectacular examples of use of television for teaching have come in developing areas like Samoa and Niger, Colombia and Peru. In those cases the problems have been so challenging that it has been possible to think boldly of new ways of solving them, and to get sufficient support behind the effort. I think it is fair to say that most American school systems have not felt quite that order of need for television. It is likewise true that most American schools which have tried television have been the schools that needed it least. Perhaps not until we turn the power of instructional television on our deprived schools (which are to be found on the mainland as well as in Samoa and the Trust Territories), or to meet the likely expansion of schooling from ages 3 to 20, with thirteenth and fourteenth grades appearing in every city of over, say, 15,000 people, will we ourselves be challenged to use the medium to its full capacity.

(2) A teaching-learning system.

Instructional television hasn't worked well when it has been used like films – at the option and at the time of particular need felt by the teacher. Perhaps some day when we have enough inexpensive videotape recorders in the schools it may be feasible to use it that way. But the way it has worked most effectively, whether in developing or developed countries, is when it is integrated into a teaching-learning system – not when it is used alone, or when it is used as an optional visual aid. This requires the teacher in the classroom to learn a new role, and to many teachers at first this role may be threatening or uncomfortable. This has been the source of most teacher resistance, where it has occurred. But experience indicates that the role need be neither threatening nor unpleasant; in fact, it may be more satisfying than the role of being the only teacher in the classroom and in complete command of everything that happens there. And once the transition period has been passed, both teacher and student attitudes toward the new way of teaching improve steadily with experience with it. The experience of working in a cooperative team of studio teacher, classroom teacher, materials expert, and others as needed, can be a most satisfying one, once the roles are learned.

I would like to digress and suggest that one of the old stereotypes about television teaching – that it is a passive experience for the pupils – is destroyed in every place where instructional television is being used effectively. The classes are extremely active. The pupils talk to the teacher on the tube, practice as directed, answer the teacher's questions, pay as close attention, and, in elementary schools at least, seem to enjoy the experience at least as much as with face-to-face instruction. The television-teaching-team is not an encouragement to passivity, but rather an opportunity to stimulate learning activity.

(3) Support.

Any innovation needs support. Instructional television, which is expensive and somewhat disruptive to old roles, and which requires obedience to central schedules, needs it especially. Governor Lee in Samoa, who was strong and determined enough to meet the resistance he encountered from certain mainland teachers who did not believe in television; Superintendent Brish who was firm without being rough; the President and other chief officials of Niger who quieted some incipient uneasiness by themselves coming to visit the experiment – this kind of support is priceless in the infant years of instructional television. And beyond that, the financial support necessary to have an adequate technical service and an adequate nucleus of skilled personnel, is a sine qua non. The magnificent technical equipment of Samoa has made a difficult task much easier. On the other hand, I could tell you about small stations where the power is forever failing, where receiving sets are not maintained, where teachers are struggling with little help, little feedback, scant experience. I do not have to tell you that there is little chance of using ITV effectively under conditions like those just cited.

(4) Size.

There is a certain critical mass beyond which instructional television seems to make an important impact and works most efficiently. Part of this, of course, comes from the economies of scale which are made possible by larger operations. Colombia, for example, can deliver television for about 5 cents per pupil hour, but it is serving a quarter of a million pupils. Some unit costs, where the system is notably under-used, run as high as $2 per pupil hour. We find that in most systems, instructional television is under-used rather than over-used. Spectacular differences in unit costs can be achieved even by balancing up the investment in programs against the investment in receiving equipment. This is something we cannot develop in this paper, but you can find details in the case studies. Beyond the financial effect, though, there is a psychological impact that comes with size. It is impossible not to take the Samoa operation seriously. It is impossible not to take Hagerstown seriously. The teachers, the students, the communities know it. Observers have reported from several places that the very fact of the existence of a considerable ITV operation has made all developmental change easier in that area.

(5) Planning.

Let us merely say that most countries and school systems have bumbled into instructional television. Many of them have wakened and found it under the Christmas tree – the equipment being intended mainly for entertainment or persuasion rather than education. Many countries that have launched into television have been overwhelmed by the sheer logistics of getting on the air. There has been all too little time to sit back and decide what the new tool might best be used for, and what related changes might be made, and what will be required outside of television programs to make the new system work. And yet there will never be a better time, than just before the introduction of a powerful new teaching-learning system like television, for a school system or a developing nation to sit back quietly for a while and consider goals and means and desirable directions of change.

There is no magic about instructional television. It is merely a pipeline into the classroom – a way to share and redistribute teaching resources. And yet it puts a wand of potential magic into the hands of any school system that has a real need to use it and the willingness to learn to use it well.

FOOTNOTE

1. The countries (or other political units) were Algiers, Australia, Colombia, Ghana, Honduras, India, Italy, Ivory Coast, Japan, New Zealand, Niger, Nigeria, Peru, Samoa, Thailand, Togo, United States, UNRWA. The case studies have now been published under the title, New educational media in action: case studies for planners. Paris: UNESCO and International Institute for Educational Planning, 1967, 3 vols. An interpretive volume has been published with them, Schramm, W., Coombs, P. H., Kahnert, F., and Lyle, J. The new media: memo to educational planners. Paris: UNESCO and International Institute for Educational Planning, 1967.

REFERENCE

Beeby, C. E. The quality of education in developing countries. Cambridge, Massachusetts: Harvard University Press, 1966.

Instructional Research: Some Aspects of Its Status,

Arthur A. Lumsdaine
University of Washington
Seattle, Washington

I WOULD LIKE to structure my paper in terms of the following main topics: (1) instructional-media research as related to research on learning and instruction more broadly; (2) classes of problems which are involved in the assessment of instructional effects, both at the applied level and at the "basic research" level; and (3) some considerations regarding the priorities of what it is most important to do in the instructional-research field at the present time. In relation to the question of priorities, I will also review the status of alleged "findings" of hopefully relevant basic research; some serious deficiencies of such "findings"; and some notions concerning requirements for improving the state of research practice.

PUTTING INSTRUCTIONAL-MEDIA RESEARCH IN CONTEXT

The first point I should like to make is one which people working in instructional-media research have realized in varying degrees for a long period of time. This is that the important problems in instructional-media research are actually broader problems of research on the effects of the management of learning and hence basically research on learning. Most of the laboratory research on learning has at least potential, even if unrealized, relevance to the design of instructional materials and procedures. However, much of it does not seem very useful when one tries to apply it to practical problems of instruction – as those who teach courses in the psychology of learning often realize when they attempt to apply their own principles.

Most of the important problems for research on basic variables in instruction--i. e., stimulus and response variables influencing the effectiveness with which instruction produces learning--have received some attention in the field of instructional-media research. In my opinion, in fact, more incisive research on instructional variables has, thus far, been done within the instructional-media field than in any other context. I think that this statement is not just a reflection of a parochial bias on my part; rather, I think it reflects the fact that the encapsulation of instructional procedures in specific media which are mechanically reproducible (like films and self-instructional programs) gives a degree of control over the variables involved in presentation, exposition, instruction, and persuasion that is very hard to achieve in other forms of research. Where a particular experimental treatment can be carefully and deliberately implemented in advance, as in a film in the cutting room, there is much less dilution from the concept, or variable, or procedure presumably being implemented, to its actual implementation in the process of instruction.

During the past few years, I have been especially concerned with what has been referred to as "programmed instruction." Like other catch phrases in the education and educational-research fields (such as "systems," "team teaching," "ungraded," etc), this is a term that I wish had never been popularized. It is not what you call something, but the way you do it that matters. Also, it has been insufficiently recognized that the work in the programmed-learning field needs to be considered as a special aspect of the broader field of instructional media (although some of the Skinnerians don't like to think of it that way). "Instructional media," in turn, is part of the still more general field of instruction, which is a special aspect of the more general field of human learning. At each level it seems desirable to go back to basic learning principles where possible, as a source (but not the only source) of hypotheses or ideas for development and test.

Not all of the good suggestions for effective instruction come as safe deductions from Hull's postulates or any other simple, single source. Such suggestions also stem partly from analysis of subject matter--e.g., in Gagné's analysis of learning sets--and partly they arise in other ways. For example, in the process of developing specific instructional products one learns a great deal about what variables seem to be important or worth looking into. One formulates hypotheses that a learning theorist, as such, would never think of when one starts to attack specific, concrete problems in education. A related point which has been made in other contexts is that it is generally unwise, in my view, for people who wish to do fruitful basic research in learning to get too far away from the actual problems of school subject matter, from the things that learners really have to learn. On the other hand, I also think that it is important, in trying to analyze the essentials of instructional-media (or other instructional) situations, to have at least a good talking vocabulary with which to interact with people whose main concern is the basic psychology of learning.

I have stressed that the instructional-media field, in its most important aspects, must relate closely to the more general field of research on instruction and learning. However, I also emphasize that there are certain features of instructional media that make them very useful as tools for research, and also objects of research (cf. Lumsdaine, 1963). Thus, the general property of stimulus reproducibility in instructional media greatly facilitates scientific study of instructional variables. A related factor is that in producing instructional-media programs the implementation of instruction is in the hands of a relatively small number of people, thereby allowing less dilution in the application of instructional principles, than when one tries to implement these same principles by influencing the behavior of thousands of classroom teachers. Finally, the ability to reproduce instructional materials widely means that application of applied research results is multiplied many fold. It is more reasonable to put out a large investment to improve the effectiveness of materials which will be used by millions of students than it is to invest comparable resources in research on procedures which are only going to be used by a much smaller number of students.

SOME GENERAL QUESTIONS FOR RESEARCH ON THE EFFECTS OF INSTRUCTION

I would like now to identify some of the questions which I think one should ask about the effects of instruction and which I think are logically prior to the formulation of specific research studies. The first question is, "What should the effects of instruction be?" Second, simply, "What are they, in any given instance?" Third, "What accounts for them?" (This is where a great deal of research emphasis has been.) Fourth, "How do you predict them?" Fifth, "How do you improve them?" And sixth, "What is the cost (and 'cost-effectiveness') of improving them--how much is worth paying for a given amount of improvement?"

WHAT EFFECTS SHOULD BE

A number of people have pointed out that in evaluating instructional programs there are several kinds of criteria used. In the report which was published by the Joint Committee on Programmed Instruction and Teaching Machines (1966) three classes of criteria were distinguished: first, appropriateness of various possible effects; second, the effectiveness with which sought-for outcomes were brought about; and third, the feasibility of using the materials, including cost. Considering just the first two, it is quite evident that one can do a very effective job of teaching the wrong thing. (It could even be held that doing an ineffective job of teaching the wrong thing may be better.) What I want to call attention to here is the far underdeveloped research field of constructive and imaginative attempts to improve on the basis for determining what ought to be taught. Very little effort has been put forth on this as compared with approaches on how to teach a given thing more effectively. This applies all the way from kindergarten to graduate work.

Skipping over the second question, consider Question Three: What accounts for the effects of instruction? Here we have perhaps attained the stage of sophistication in which we turn away from unanswerable questions (such as, Which are better, films or lectures? or, Is conventional instruction better than television?), and have moved to questions of what variables account for particular effects. That is, we have become better aware of the consequences of the obvious fact that a good film will always beat a poor lecture and vice versa, and of the fact that such terms as "conventional instruction" have been so poorly defined as to be almost completely ungeneralizable.

Question Three gets at the attempt to manipulate variables in order to see which kinds of stimulus-input variables make a difference in instructional outcomes. Obviously, if this type of research were to be successful in building an applicable science of learning, within the confines of a given set of instructional objectives, we would then be able to better answer Question Four, (How do you predict the effects of instruction?) as well as Question Five, (How do you improve them?). That is, we would have a basis for evaluating an instructional program of any sort by looking at it, analyzing its properties, seeing whether it conformed to characteristics of instruction which have been demonstrated to be more effective than other forms of instruction, and thus would be able to evaluate, a priori, any particular piece of instructional material, or program, as to its effectiveness.

Some of my colleagues feel that this can now be done with respect to programmed instruction. In my opinion, they are quite mistaken. For example, in one of Rothkopf's (1963) studies, (admittedly one which is too limited to be definitive) a group of "experts" (students who had recently been through a course in the "proper" way to construct instructional programs) were asked to rate seven programs. These seven programs all dealt with the same subject matter and, without the knowledge of the judges, their actual effectiveness in teaching specified kinds of outcomes had been previously determined by trying each of the programs on seven randomly selected groups of students. When rated effectiveness in that set of seven programs was correlated with measured effectiveness (as measured by what students actually learned), Rothkopf concluded that the instructional effects were indeed "predictable": the correlation was minus .75. All that need be done was to get judges to rate them and then do exactly the opposite. Such data support the contention that the effects of instruction are not really very well predictable on the basis merely of inspecting the features of the instructional materials and deciding whether they conform to the implications of a "science of learning" or to the accumulated wisdom of teachers. This means that either we are applying the wrong findings or we have a great deal more to find out about the effective variables in instruction. This in turn takes us back to Question Two, What are the effects of instruction?

I contend that we might well set for ourselves the perhaps modest objective of finding out what we are actually doing in much on-going instruction. I think we at least need to know what effects we are producing before putting such exclusive concentration of effort into the more elegant question of what specific variables account for these effects. In many of the studies of instructional effects, one problem has been the lack of significant differences, reported with dismal uniformity. This lack would have been predictable if the investigators had bothered to run an initial test of any one treatment of the instructional materials (preferably the best), using available criterion measures of output. Often this would have shown them that, in fact, the whole program was teaching at such a low level of efficiency that the likelihood of revealing any differences among the treatments was indeed very poor. All one has to do is to read the instructional research literature to find many instances of this kind, and I must even admit that this has been true of some of the studies which I have been responsible for sponsoring. It is a very simple truth, but one which has eluded investigators for a long time, that if we are going to investigate factors that make instruction effective we have to get the instruction effective in the first place and then find out something about how it got that way, instead of working hard at manipulating variables in a piece of instruction which we later find out is operating at a low level of efficiency. One of the implications of this is that the classical sequence of research and (then) development often must be reversed. In other words, one must do the development first; only when something effective has been developed, do we then have an adequate tool with which to conduct useful research. This is an oversimplification, of course, but it remains a crucial and an all-too-often ignored consideration in the instructional research currently being done (and continuing to be supported).

Regarding Questions Four and Five (How to predict the effects of instruction and how to improve them), it may be noted that both imply important kinds of purposes. However, it can be argued that prediction is of no use in itself unless it can ultimately be used to bring about better instruction. One way to do this is through better selection of existing materials (as well as making better ones).

The conduct of adequate empirical evaluation studies for this purpose, which would require the detailed testing of every piece of marketed instructional material, is an extremely time-consuming business. It would obviously be a great advance if one could reliably predict which materials would be more effective than others, as a basis for selecting from available materials within a given set of instructional purposes. In any case, the ultimate applied purpose of fundamental knowledge about the influence of instructional variables on learning outcomes is to be able to know how to improve the instructional product.

A vexing problem of instructional research for this purpose concerns the extent to which the independent variables named by the investigator are idiosyncratic to him. For example, in experiments on "student participation" or "varied repetition," (or you name it), the independent variable is seldom defined in terms sufficiently operational to provide reasonable assurance that someone else would be doing essentially the same thing if he said he were studying the nominally "same variable." That is, for most of the pedagogical variables derived from either common teaching practice or learning theory, the specific way in which the "variable" is implemented can turn out to be a great deal more important than what variable it nominally is. We, therefore, have findings in which the "same variable" (e.g., overtness of response) turns out to have one effect when explored by one investigator and another effect when explored by another investigator. The conclusion would seem to be that, pending such time as we are better able to define what our instructional variables mean in reproducible, operational terms, we had better assume the burden of demonstrating explicitly the generality of our findings. The deduction from this is that in an instructional experiment in which one manipulates a named variable or procedure, the critical number of "degrees of freedom" on which generalization forms the experiment depends is a function of the number of experimenters (or programmers), rather than the number of subjects (learners). Put another way: if we are really going to say that we have found a dependable rule or generalization about an instructional procedure, we have got to show that we can write a set of instructions which will allow another experimenter, programmer, or teacher to create substantially similar effects to those we obtained. I know of only one or two studies (still underway) in which this is systematically being done. In one of these studies the number of instructional vehicles representing each treatment are written by six different programmers following a given instruction, as against six other programmers who are following a different instruction. The number of degrees of freedom in that analysis will be something in the order of either 10, 12, 5, or 6, depending upon what kind of statistical test is used, but it will not be a function of the number of students or classes that are taught.

This very important point can be summarized simply by saying that we have to be able to show, in a scientifically acceptable fashion, that someone else can manipulate and implement the variable under consideration. At present, I know of no way to do this, for most variables at least, except by brute empirical demonstration. This is one of my main reasons for the quotes around the word "findings"; for, of the experimental results on instruction, very few, if any, have been systematically replicated in a way that provides this kind of formal demonstration of their generalizability outside of the particular program or programmer responsible.

To consider Question Six (concerning cost and cost effectiveness) requires that we first consider basic objectives. In terms of the kinds of outcomes we are trying to achieve, we have ultimately got to translate progress toward these objectives into something that is cost accountable. I think the things that have the best chance of becoming cost accountable are things that are correlates of time, rather than being correlates of achievement level. For this reason and other, I think a strong case can be made for Skinner's contention that, in effect, we should use as our dependent variable the time to achieve a given minimum criterion of proficiency, rather than the levels of proficiency attained after a fixed number of hours of instruction as shown by the resulting distribution of scores.

A related problem is that our realistic goals in education are often how to bring people up to a point where they can quickly relearn rather than to maximize how much is retained by the end of the semester. There is simply getting to be just too much to know and too much to learn to keep it all on tap, and part of the problem of education is, increasingly, going to be to determine how much we ought to get "kept" in a man's head versus how much we ought to help him out on the job by better information sources or refresher training. We might ask how much a realistic criterion of the goodness of instruction is the speed with which a person can get back up to proficiency, rather than how long it takes him to get there in the first place.

A difficulty with this kind of criterion (aside from the break with tradition that it implies) is that it greatly increases the cost and difficulty of experimentation. If one is going to measure relearning of paired associates, the savings method of the psychological laboratories is a rather easy method to apply. But if one is concerned with the learning of a subject matter like calculus, instead of pairs of nonsense syllables, then the question arises: What is the material from which one does the relearning? I think we have to work hard on solving such problems, because time expended in order to attain or regain acceptable proficiency offers the best promise of being translatable into cost terms. We can form some estimate of the cost of the student time and can likewise estimate the cost of the heating for the school, the teacher's time, etc. But how to estimate the cost of the difference between 700 and 750 on a set of achievement test norms seems much more puzzling, and possibly involves the wrong kind of educational objective anyhow.

RESEARCH AND DEVELOPMENT PRIORITIES

QUANTITATIVE PRODUCT-EVALUATION DATA

I believe that a great deal more work is needed on the simple, modest question of assessing or measuring the effects of available and reproducible instructional programs; I think this perhaps unglamorous area rates a very much higher priority than it is getting. We need to know what we are accomplishing, and by and large we don't know that. The programmed-instruction people have banged the drum for data accompanying programs, and we have even set up some standards so that we could have some hope that a sort of Gresham's Law wouldn't operate with bad data driving out good data. Yet at present I know of only a very few programs on which any reasonably adequate data have been published. Somewhat more than that number of programs have something which purports to be data, but which are merely records of what a few students did as they went through some version (not necessarily the published one) of the self-instructional program. For the present I would allocate a much larger share of research budget to support lowly, "dirty," applied work on getting product specifications for available programs that potentially will have wide use, and at least relatively less of it to what are, all too often I think, futile attempts to do fundamental research. (I am well aware that some of my colleagues will regard this as utter heresy.)

THE PROBLEM OF WHAT TO TEACH

Another act of high research priorities is related to Question One, - What should the effects of instruction be? We have an enormous range of opinion on this. We also have an increasing amount of knowledge to be taught and yet no rational basis for selecting what we should teach, not only to graduate students in specialized areas and to undergraduates who are acquiring a "liberal" education (presumably to prepare them in some fashion to deal with the problems of the world), but even, to a considerable extent, to the elementary-school children. The problem does get somewhat easier at the lower grade levels, as there are certain fundamental skills which appear necessary; but beyond the elementary-school grades I think we are flying blind in terms of what we are teaching. A very good case can be made, for example, for teaching high-school students the fundamentals of sampling and scientific inference rather than Euclidean geometry.

In any case, I believe it would be very helpful to try to do more extensive survey research to attempt to identify areas where agreement and consensus can be reached as to what should be taught. At this time a very promising attempt of this kind is being made by a group under the leadership of Ralph Tyler (1966).

I also feel that some experimental evidence could be quite useful in sharpening up the question of what to teach. Many of the questions of what ought to be taught are ultimately transfer-of-training questions. If a more sustained attempt were made to define and characterize the kinds of educational outcomes we really want, then we could try to verify some of the working assumptions which we seem to make as to what kinds of learning transfer to these outcomes. We would be then moving in the direction of an experimental attack on what ought to be taught. There is, of course, so much hypocrisy in the whole field anyhow that even to get people to be a little more honest would be helpful. ("The purpose of my course is to teach the students to think"; but the final examination consists of multiple-choice questions about the facts in the subject matter.) Very few determined attempts have been made to formulate instructional course objectives so as to reflect in the final examination, as well as in the development of the course, what the professor professes is really the most important objective of his educational endeavor. Some hopeful steps in this direction have been taken. (The last course I taught got almost to first base in this respect, and the next time I teach it I hope I can get a little further along.)

EMPIRICALLY GUIDED PRODUCT DEVELOPMENT

Next to a determined attempt at research on what should be taught, I would say the next most important priority within the total research and development sphere is not in the research field proper but in the area of development. In my opinion, a great deal more support than is presently the case should be devoted to empirically guided development projects. (The U.S. Office of Education has not as yet fully taken my advice on this point.) Perhaps one way to proceed would be to require product development as one element of an R and D package. That is to say, if one is to conduct an experiment which purports to "test a hypothesis" and derive some "fundamental knowledge," and if one is realistically pessimistic about the likelihood that he will succeed in this objective, perhaps he might be asked to have a second string to his bow. If, for example, all such experiments were conducted with actual subject matters taught in schools and did both of these two things, then if they failed in what was the nearest and dearest objective to the heart of the investigator, they still would have accomplished something useful. That is, if one produced a useful package of demonstrably reasonably effective new instructional materials that were needed as a vehicle for the experimentation, and if this product of the research were to be made available to some appropriate distribution agency, then the sole product would not be confined to a dubious conclusion in the pages of a journal; the product would include an actual on-the-shelf instructional item that could be used. If in addition a little illumination was shed on the basic dynamics of instructional variables, so much the better. If it weren't, at least one would have a useful film, self-instructional program, book, etc. which would help to justify the investment that went into the research project.

Also, some studies could usefully take an existing product, widely available and in current use, and use it in a research project as the vehicle for making experimental manipulations to inquire about the effects of instructional variables. Then, data giving a good profile on the actual teaching accomplishments of this product could be made available in the form of product-specification data. This would thus provide a well-defined set of measures, analyzed item by item, so that one would know which points of information, skill, knowledge, or other behavioral outcome were increased and by how much by the use of the basic product in a specified manner with specified kinds of learners. One would then at least have some useful product data, regardless of whether or not his attempt at scientific inquiry on instructional variables were successful.

In summary, I would tend to plug first for empirically-guided development studies, second for other evaluative studies, and (a low third) for purely "basic" research, given the present state of the art. A good way to have one's cake and eat it too here is to do studies which combine investigation of instructional dynamics with the development or assessment of a product, in such a way as to create either a useful product or useful product data, in addition to whatever illumination may be shed on more fundamental, generalizable problems.

SOME WEAKNESSES IN "BASIC" RESEARCH "FINDINGS"

As to the frequency or likelihood of genuine illumination, suppose that one applies the criteria any real scientist has the right and obligation to apply to research findings, including that of replication; and that one throws out all of the findings which lack any adequate guarantee of the sensitivity of measuring instruments or fail to show sophistication about power of statistical tests used, and that one throws out all of those indefensible findings which have proven the impossible by violating the injunction against proving the null hypothesis. One then gets down to the relatively few findings in which significant differences are reported, of which he will still find that very few have been replicated. Of the small number of these, perhaps most are in the field of the role of active response of students. Even here, however, it is difficult to say much more than simply that some forms of student responding under certain circumstances can prove quite helpful; the conditions under which one form or another of responding can help how much are still far from clear. Also, there seems to be some pretty consistent evidence that we have probably underestimated the extent to which we ought to help the student out by prompting; but that evidence is more secure for paired-associate learning than for more involved kinds of instructional materials.

A more gross generalization that I think could be supported as a kind of by-product is that most instructional lessons, films, television lessons, and college courses try to teach more than they possibly can in the time that is allotted them, at least on the frequent assumption that these are the sole medium of instruction. Demonstrably, almost every instructional program as initially presented is found to need to be extended, to need to be given more time (or else its objectives need to be reduced to more modest size) when the program is tested to reveal a profile of what students actually learn. This can be found in study after study when the results have been reported in such a way that you can find out what was actually learned, not just in terms of P-value or a misleading dichotomy between "significant" and nonsignificant findings.

The conclusion may be a discouraging one, but the most important variable in the effectiveness of instruction all too often seems to be simply the amount of time which is spent in instruction. Thus, if one wishes to establish the importance of a procedure one has to do a lot more than has customarily been done in ruling out not only the null hypothesis but also what some of us have come jokingly to call the "erg hypothesis." Too often the procedures which are shown to be better just happen to be also the procedures that take more time, and with no sensible way of refuting the skeptic who says "So what? Naturally, if you spend more time you will teach more."

Many of the weaknesses of the instructional research to date are, I think, due to the lack of the promulgation and enforcement, for so-called basic research studies, of adequate standards of scientific reporting (which we have tried to formulate for evaluation studies in the Joint Committee report mentioned earlier). The generally poor standards of reporting allow the ambiguous use of terms, and failures to distinguish interpretations from actual findings, to a shocking degree in much of the educational research literature. It is almost impossible to read through a set of papers in this literature without finding, upon examination, that many conclusions stated as if they follow from the data are really just some speculations produced by the experimenter to explain what he found or why he didn't get what he expected to find. I am exaggerating only slightly. It is an agonizingly difficult task to attempt (e.g., Lumsdaine and May, 1965) to glean from research literature in the instructional-media field that it is that has actually been demonstrated. In many instances one is unable to get the basic data necessary to find out whether the findings warrant reporting, yet perusual of the available reports makes it clear that without going back to the basic data one cannot determine what conclusions, if any, the available data justify. One simply cannot believe, by and large, what one reads in abstracts, secondary sources or even in primary research reports. And if one tries to check on the latter, he often finds that the basic data are inaccessible.

The APA-AERA Joint Committee recommended, for evaluation studies, that the basic data should be preserved, preferably in the form of a complete matrix of each student's response to each test item, so that the summary results can be verified. The value of such data preservation is twofold: (a) very often serious errors are made in the data summarization; there are nothing like the uniform standards of checking in educational research data that one expects in physics or in the experimental psychology laboratory; and (b) very often the findings, even if accurate, are recorded in a summary form that precludes answering relevant additional questions to which the original report was not addressed. By way of example, there is a very good a priori reason and considerable empirical support for believing that the effects of different kinds of instruction are likely to interact significantly with intellectual ability, as measured at least grossly by IQ, Army General Classification Test scores, etc. Yet study after study appears in the literature in which only the overall data are reported, with no indication whatever, even though the basic data contains this information, of how the results interacted with, or were modified by, differences in student ability. Even when differential analyses are a function of ability are made it is often impossible, without going back to the basic data, for one study to be compared with another which has, let us say, made one split on an ability measure while another study has made a different split, or used a different measure. It would be perfectly possible, and often of considerable interest, to see if these studies indeed contradict or confirm each other, but a person cannot do it because he doesn't have the same split on the differentiating variable. This often would be very easy to rectify if the basic data were preserved, as also would be true in other instances in which the analysis has been incorrectly performed but might be corrected by going back to the basic data.

Clearly I believe there is great need for general improvement in reporting standards. As another example, one of the most common errors that we try to teach introductory students to avoid is that special form of the unprovability of the null hypothesis which is exemplified in the following argument: Group A and B did not differ significantly before getting treatment X; they differed significantly after the treatment, and therefore the effect of the treatment was significantly different for the two groups. Of course this argument is quite invalid; what can be shown rigourously is only that the difference between before and after was greater for one group than for another. It is quite possible to document the proposition that treatment X made more difference for group A than for group B but, again, the basic data by which this might be tested (e.g. by a second-order t-test) have not been preserved so as to be available for making the test.

SUMMARY

In summary, I suspect that the most important thing that can be done to improve instructional research methodologically is simply to try to clean up the standards of reporting and thereby the standards of performance for educational experiments--at the semantic level, at the data-preservation level, and at the level of insisting on a more complete description of what was actually done in the experiment. If this is not done, and here I utter a Cassandra-like note, I believe ultimately we are going to kill the goose that is laying all the Research and Development Center eggs. We need desparately to clean up our experiments methodologically, and to insist on at least the same standards of rigor for educational research reports and journal articles that are applied by the editors of the Journal of Experimental Psychology.

I should like to add that the dependent-variable measures we are using in educational experiments are also in great need of improvement. For instance, there are few studies available in which the effects of instruction or indoctrination on "attitudes" have transcended the level of mere verbal statements on a questionnaire by the respondents as to how they think they were affected or how they believe they would behave in a particular situation. Only a few studies have gone beyond this. One such study was conducted by Mark May at Yale some years ago (May and Jenkinson, 1958). Instead of simply asking the students, May observed how students actually behaved so as to obtain a measure of motivational influence of two alternative forms of a film. George Gropper and I did a somewhat similar study a few years ago in which we assessed the effects of some television broadcasts in alternative forms presented over WQED in Pittsburgh (Gropper and Lumsdaine, 1961). Janis and his associates at Yale have also used observation of real criterion behaviors, instead of mere verbal surrogates, in studying the effects of various forms of communications (see Janis and Feshback, 1953).

But such studies represent only a bare start in the direction of what needs to be done. That this is so leads to a final conclusion I would like to draw: namely, I think it would be much better if we would pick out a relatively small number of problems, and a relatively small number of standard subject matters, for such experimentation as purports to get into the real dynamics of instructional variables, and stick to them and really try to pin a few things down by extensive varied replications with good dependent measures, instead of striving so much after novelty that we hop from problem to problem with a series of superficial studies that cannot be compared with each other and that really prove nothing, individually as well as collectively. I believe this can be said, as a matter of fact, throughout the whole spectrum of educational-psychological research. There are those who advocate, and a few who even put into practice, the insistence that Master's theses should always be replications of experiments rather than attempts at new experiments. I personally think that this would be a healthy general practice; for instructional research I know it would be. We need a great deal more replication and elaboration of a few putative findings that look fairly promising, and much less uncoordinated striving after novelty in an attempt to win kudos in "pioneering."

REFERENCES

Gropper, G. L., and Lumsdaine, A. A. The use of student response to improve televised instruction: An overview. Report No. 7, Studies in televised instruction. Report No. AIR-C13-61-FR-245 (VII). Pittsburgh: American Institute for Research, 1961.

Janis, J. L. and Feshback, S. Effects of fear-arousing communications. Journal of Abnormal and Social Psychology, 1953, 48, 78-92.

Joint Committee on Programmed Instruction and Teaching Machines. Recommendations for reporting the effectiveness of programmed instructional materials. Washington D. C.: National Education Association, 1966.

Lumsdaine, A. A. Instruments and media of instruction. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 583-682.

Lumsdaine, A. A. and May, M. A. Mass communication and educational media. Annual Review of Psychology, 1965, 16, 475-534.

May, M. A., and Jenkinson, Nelle L. The reading of a book based on a film. In M. A. May and A. A. Lumsdaine, Learning from films. New Haven, Connecticut: Yale University Press, 1958. Pp. 195-203.

Rothkopf, E. Z. Some observations on predicting instructional effectiveness by simple inspection. Journal of Programmed Instruction, 1963, 2, 19-20.

Tyler, R. W. Let's clear the air on assessing education. Nation's Schools, 1966, 77(2), 68-70.

Defects, and Needs

SOCRATES, A Computer-Based Instructional System in Theory and Research

Lawrence M. Stolurow
Harvard Computing Center
Cambridge, Massachusetts

THIS PAPER describes a cybernetic computer-based instructional system, SOCRATES, the teaching model which led to its development, and some of the research accomplished with it. Early work on non-cybernetic instructional systems is described in Coulson (1962), Silberman and Coulson (1962), and Bushnell (1962). For a description of other existing computer assisted instructional (CAI) systems see Stolurow and Davis (1965), Gentile (1965), Hansen (1966), and Zinn (1965).

SOCRATES

The acronym, SOCRATES, is System for Organizing Content to Review And Teach Educational Subjects. It consists of a group of student input-output (I/O) stations wired to a digital computer through a relay rack. It is a computer-based, multiple access, instructional system designed to meet the requirements of an idiographic contingency model of the instructional process (Stolurow, 1961; 1965b, 1965c). The model defines the presumably critical categories of instruction and treats these as system functions. The system permits controlled implementation of the model as well as the collection of data which could lead to its further development.

Figure 1 shows the configuration of the system, called SOCRATES II. It consisted of an IBM 1620 computer with a 1311 disk for additional storage, plus a 1710 Multiplex and terminal unit linked to several student I/O stations, or consoles.

Figure 2 shows some of the individual student I/O stations, each of which displayed 35 mm frames of a film strip with a 1,500 frame capacity and provided the learner with fifteen keys for either single (selected) or multiple (constructed) response to a display (Stolurow, 1966).

AN IDIOGRAPHIC CONTINGENCY MODEL OF TEACHING

The model of the teaching-learning process is an idiographic contingency conception (Stolurow, 1965a, 1965b). It is based on the assumption that individualized instruction is a multiple stage decision process that is diagnostic, prescriptive and corrective in its use of psychological information to make decisions about teaching. While the decision process is basically continuous three stages can be identified. In the first stage the pre-tutorial decisions are made; in the second, the tutorial decisions are made; and in the third, the program revisions are made. In the first two stages there are two levels of operation which make the system highly adaptive to individual differences.

Pre-tutorial Decisions

In the pre-tutorial stage, one set of decisions determines the assignment of each student to an instructional station; the other set determines the educational program to be used in teaching a student.

Automatic assignment. When a student entered the instructional area, he could sit at any vacant student station each of which contained a MASTER I/O unit. The MASTER I/O unit, connected to the computer, contained a library of materials on film

Figure 1. SOCRATES II

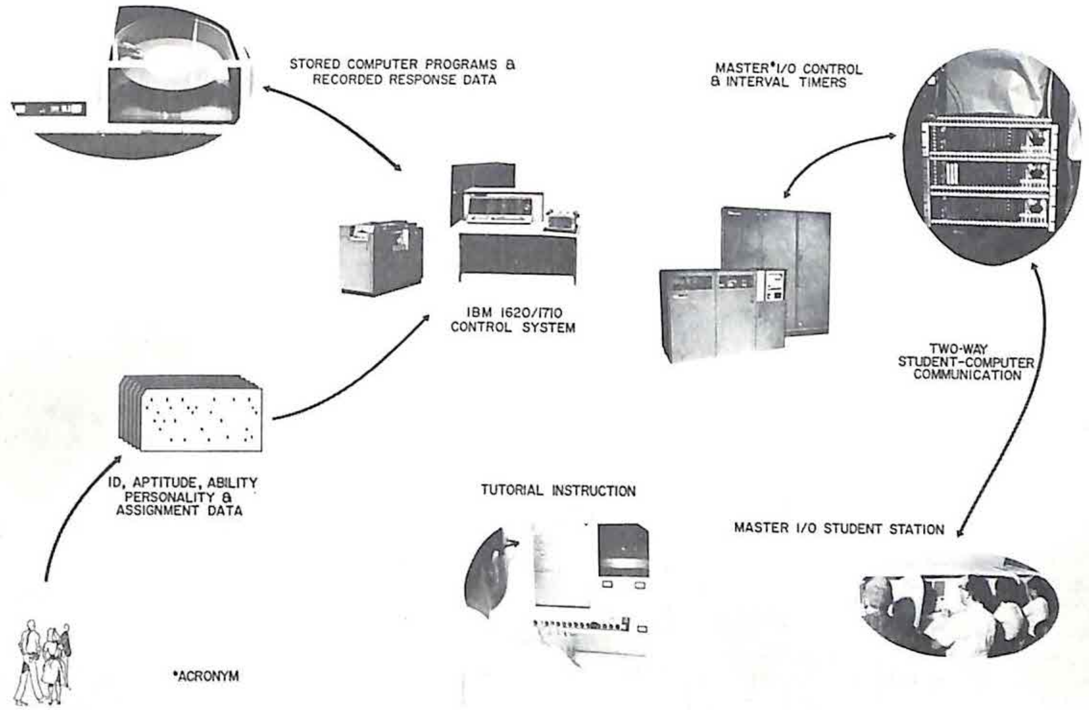
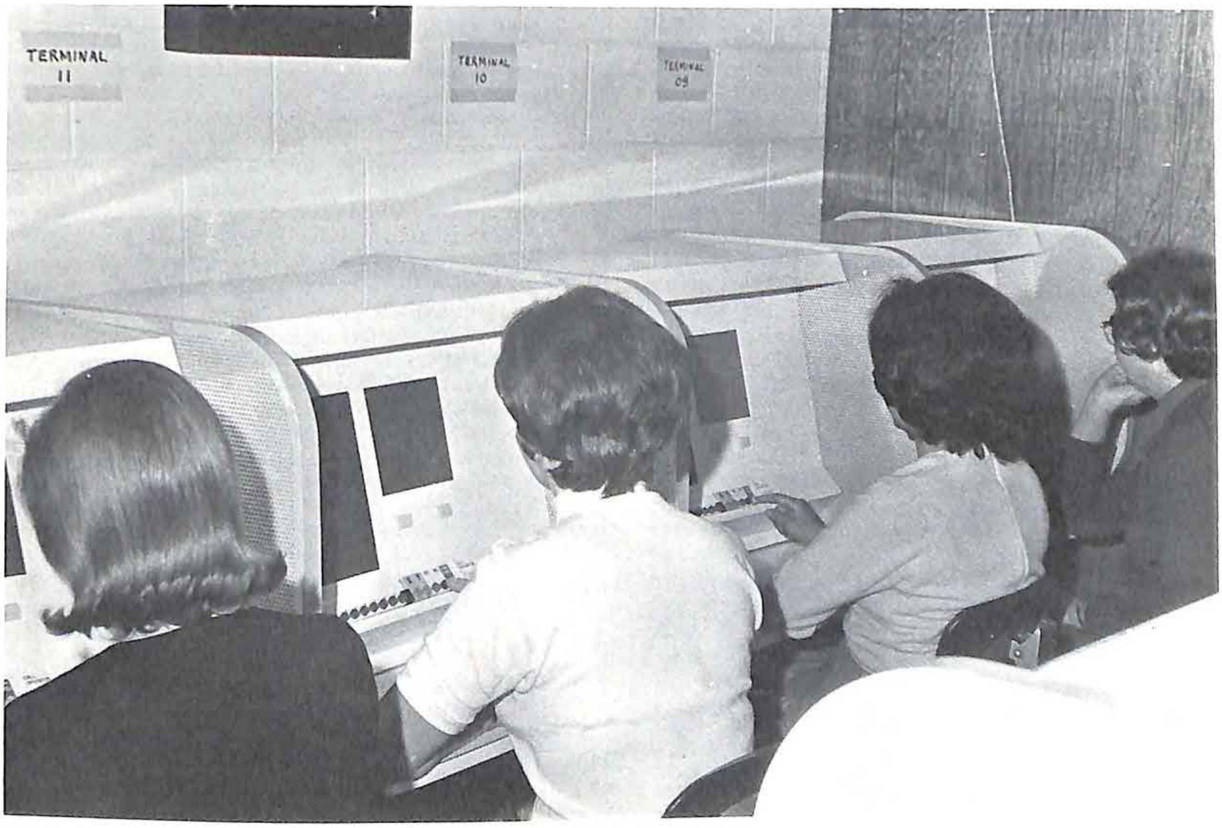


Figure 2. Individual student I/O stations



(1,500 frames) and each could contain a different library (film strip) if this seemed feasible and desirable. Feasibility depends, in part, upon the amount of core storage one needs for the computer programs used to process each of the educational programs contained in a library.

The purpose of the automatic assignment was to simplify and expedite the process of getting the student on-line with the computer. The procedure in automatic assignment consisted of the following. A student went to any station – for example, 11. If it was not one at which the instructional materials needed were contained, he had to be directed to another station which had the appropriate material. In order for the system to do this automatically, the student read an initial frame containing instructions to both enter his ID number using the keyboard and to press the "I am thru" key. When he did this, another frame appeared on which a film number was displayed. Every film contained a unique number. After the student entered his ID, he entered the film number. The system then compared the number with the one it had stored as the appropriate assignment for that student. If there was a match, instruction began at that station. If they did not match, however, the system searched all available stations containing the film which the student should see and advised him of the one available; it instructed him to go there. This was done by showing him a station number. Once the student was seated at an appropriate station, the pre-tutorial phase of instruction began.

Initial selection of an individual's program. In order for the system to make pre-tutorial decisions about the student's instructional program, it is necessary to have a computer program which processes information about the educational program and about the student. The objectives of the instructional program must be stated in behavioral terms, namely, the type and level of performance that is required. Each objective is a "do statement"; it contains a relevant verb and a description of content and conditions under which performance is to occur. Each film included frames containing: (a) information; (b) directions (e.g., procedure); (c) questions; (d) response options; (e) knowledge of results; and (f) evaluative feedback. Performance requirement for correct response, expected wrong responses, and the maximum training time that can be used are all specified. The student's pretest scores indicate his beginning knowledge in relation to each objective, his aptitudes, and his personality characteristics.

More specifically, whenever any of the student's pretest scores – his entry level of performance – equals or exceeds the minimum terminal level, the set of frames relating to that objective is eliminated for that student. The remaining set of objectives has to be taught; to do this, the system needs to be programmed to make the initial decisions about the rules to be used in teaching the remaining set of objectives. The set of decisions used in teaching is its instructional logic. When the logic is expressed as a set of rules for interaction with a student or, in other words, when the contingencies to be used with a student are decided upon, then the teaching strategy is defined and the program specified. The initial decision is to identify the instructional logic. The flow diagram for program selection during the pre-tutorial process is presented in Stolurow (1967).

Selection of contingencies. To use data about the student in making the decision about how to teach him, it is necessary to have information about the typical, or expected, relationship between pretest scores, on the one hand, and the effects of specific treatments such as a type of feedback or a sequence, on the other hand. Each of these relationships is a possible contingency that has to be implemented by the system. The desired contingencies are expressed as "if...then" statements. To illustrate: A contingency, as a rule to be used during learning, states that if a student with low math aptitude answers correctly, then he is sent on to the next frame, but if he answers incorrectly, then he is sent back to the original material and must select another alternative. This is a correction contingency. Another example is the aptitude-sequence contingency, namely, if the student's aptitude scores are X, then the sequence of instructional material is 1, 2, 3. For example, if a student with high language aptitude correctly answers the first five questions about a passage just read, then the remaining three questions in the set are omitted and he is given new materials that he would not otherwise see. However, if he does not answer correctly, the remaining questions in that set are given. The overall instructional logic is what determines the order in which the student would receive new materials during learning. If research data show that students above a particular level of mathematical aptitude learn to add fractions better through the RULEG sequence than they do through the EGRUL sequence (Taber, et al., 1965), then a contingency between aptitude level and the RULEG sequence would be programmed and used in making decisions about teaching the addition of fractions. Students with scores outside that range might be taught with another sequencing rule, e.g., asynchrony (Detambel and Stolurow, 1956), or class-descriptive cueing (Wulff and Stolurow, 1957).

To specify another contingency each student's personality test scores can be used to determine the form of evaluative feedback informing him of the correctness of his response (Frase, 1963; Parisi, 1965). These are just two examples of other contingencies that can be used in making teaching decisions. Ideally each contingency should be based upon data from research revealing that it contributes to the effectiveness of instruction. However, if appropriate data are not available, a set of theoretical possibilities can be used to specify the type of feedback, the frame sequence, etc. Whenever the teaching system uses contingencies, it branches the course of instruction at the decision points. A system can make a decision for each student at any point in teaching and alternative contingencies can be compared in their effectiveness to determine the value of branching. In doing this, the system is being used to test hypotheses about contingencies. Thus research needs have to be satisfied as well as those of application. All rules specifying the initial set of contingencies have the status of hypotheses and, as they are used, the system acquires the necessary data relating to their utility. This also means that the system provides a means for identifying additional contingencies. The instructional logic, expressed as a set of rules, is the set of contingencies used in teaching a student using a particular set of materials. In individualized instruction the logic may be quite different for different students. The system must be programmed to select a strategy to go with the topics to be taught each student. The combination of content topics and strategy makes up an instructional program for a student.

Automatic adjustment. The system must also be able to handle the student for whom no progress could be expected to produce the minimum level of performance specified by the instructor.

Whenever the student's data indicate that he would probably not reach the instructional objectives by any available program, the system can a) teach the student some things to remove his deficiencies; b) lower its minimum acceptance criteria for knowledge; c) lower the terminal objectives; or d) reject the student. Each of these alternatives can be programmed so that it is considered automatically in terms of a predesignated priority rule. For instance, if the decision is to improve the student's proficiency, then he would be given remedial instruction in one or more of the prerequisites. For example, the student could be assigned and given an algebra program to build up his entry behavior level before he begins studying a statistics program. When he completed the algebra program, he would be tested and his new entry level of performance would be substituted for his original level and another decision would be made about teaching him.

Another alternative is to change the minimum acceptance criteria. The student below minimum-performance level on the prerequisite knowledge test but high in relevant aptitudes can be given an instructional program which is most likely to increase his final performance on the test of terminal behavior. Doing this would increase the proportion of failures, but it also would increase the absolute number of successes. Therefore, this is a decision strategy that probably would be used only if it was very important to meet a quota of trained people.

As a third alternative, the system can modify terminal objectives when it fails to find a suitable program. One way to modify objectives is to increase the maximum allowable instructional time. For each change in that parameter, the system can re-estimate a student's final performance to see if the next estimate meets the minimum terminal proficiency level. Once the estimated score equals or exceeds the minimum, the system can print this out so that the instructor can decide what to do next or it can compare the estimated time value with a new maximum value. The result of the comparison determines whether a program will be used with the student. A second way to modify objectives is to eliminate topics that have low priority. A third way is to lower the minimum level of performance which is acceptable at the end of instruction. Each of these ways of modifying objectives can be programmed so that the appropriate parameters are considered sequentially and a new estimate of final performance obtained. The system can store the estimates from each of these modifications and select that one which provided the maximum final performance estimate.

The system can be programmed to operate at two different levels in making pre-tutorial decisions. At the lower level, it makes decisions based upon a fixed set of rules and a fixed set of values for the critical parameters used by those rules. At the higher level, the system changes the values of parameters according to a different type of rule, one which determines the priority of the parametric changes and then re-evaluates the student. The priority rule itself can be varied from student to student to achieve further individualization. For example, the priority rule can be made to depend upon the relative size of the discrepancies between the original criterion values and each student's scores.

The outcome of the pre-tutorial processing is a tutorial program for each student based upon data about his aptitudes, abilities, and personality.

Teaching Program

A teaching program is made up of two parts: its logic and its content, the particular topic or set of topics to be taught. The complete set of rules used in making all of the decisions in teaching a student is the logic of his instruction. The content depends upon the student's pretest performance on a test covering the behavioral objectives of the instruction and his performance while learning. Selection from the library depends upon pretest performance. If pretest performance on a topic is already equal to the minimum acceptable final performance level, then the material used to teach it is eliminated from the student's program. Thus, the content of the student's program depends upon a contingency. His program contains only those materials which are necessary to teach the knowledge and skills which his initial level of performance indicated were inadequate in comparison with the terminal standards.

The functions involved in individualizing an instructional program are described in Stolurow and Davis (1965). Of those represented the more critical functions are the comparison of the student's responses with those expected (performance standards), the selection of particular displays, and the recording of a response history for on-line use as well as for course revision.

Teaching Strategy

A teaching strategy can be described in terms of two types of rules used in presenting a set of frames. The first type of rule determines sequence, or order, i. e., organization of the frames relating to an objective. The order in which instructional materials are presented to the student is determined by contingency rules relating to aptitude patterns. Each element of instruction is presented as a frame which requires a response (Markle, 1964). There is a variety of frames, each of which has a different purpose. The order in which they are arranged can make a difference in rate of learning and in retention (Stolurow and Lippert, 1964). Consequently, the type of frame and the sequence in which frames are arranged represent important specifications for a program. The second type of rule specifies the conditions under which the frames are presented; for example, a picture which is part of a frame may be timed, or the medium used to encode the frame might be a film, for some, or an audio tape, for others. A complete set of rules includes both types, and the set employed in teaching a student to reach an objective is referred to as a teaching strategy. For example, a young student with high mathematics aptitude might be taught with a RULEG sequence under the conditions of the correction procedure and with visually displayed frames. An older student who is also high in mathematics aptitude might be taught the same concept with a RULEG sequence under the conditions of the non-correction procedure with audio tape. Each would be taught by a different teaching strategy.

Data for instructional decisions. In order for the teaching program to be effective, the system must use appropriate data correctly. The kinds of data to be used are specified by the teaching strategy, or rules expressed as contingency statements. To make individual predictions regarding the final performance of each student from his performance history, both the data and the rule possibilities have to be represented in the computer memory and used by the control program. In this respect the system operates in the tutorial mode just as it did in the pre-tutorial mode.

If the predictions about final performance are below the desired level, the system should be capable of changing the rules it is using and the basis upon which the rules are selected. For example, regarding the latter point, the system after teaching several students using a particular teaching logic, should use the data they provided it to re-analyze and compute new limits.

Whenever the student's performance does not meet the specified standards, adjustments are made in some part of his projected program. Also the data contribute to the adjustments made in the decision rules themselves for future applications to other students. In this way his data contribute to the pool of information used in making decisions about teaching subsequent students (e.g., Smallwood, 1962; Lippert, 1967). As a result of its previous teaching, the system learns how to teach more effectively.

Two Levels of Tutorial Instruction

Two levels of tutorial instruction are possible. The lower level obtains when the system is programmed to teach with a fixed logic and a fixed content, e.g., a linear program (Skinner, 1954), drill and practice (Atkinson, 1967). The higher level obtains when the system is programmed to alter either its logic, its content, or both, for subsequent parts of the instructional program depending upon the student's performance on the earlier parts. An example of the lower level is using the logic of a linear program (Skinner, 1954); another is using the logic of an intrinsic program (Crowder, 1960). For both of these strategies the content and logic must be specified in advance. At this level all of the instructional possibilities are pre-specified including the types of feedback given all students. At this lower level of tutorial instruction a CAI system provides a pre-specified learning environment which in one case is predetermined (linear program) and in the other case is contingent on the previous response.

A distinction can be made between a pre-specified learning environment and a learning environment that unfolds. In the former, the actual displays seen by the student are specified in kind and in sequence. If the student makes response "A," then he sees display α, etc. The system operates on the basis of conditionals that are of the form "If this response, then that stimulus display." All the conditionals are specified, and all students who make response "A" see display α. In the learning environment that unfolds, all students who make response "A" do not see display α, only some of them do. What display a student sees is determined by more than one rule, so that display α, for example, could be seen by some students who made response "A" but not by others who also made that response. An example of unfolding might be to show some students a frame containing the steps in the long division algorithm worked out with the values 360 divided by 12, while others may see the steps of this algorithm worked out with 480 divided by 52; some others may see examples with decimals. Still others may see the example worked out in a completely different sequence, e.g., successive subtraction and tallying. To do this on some basis other than chance implies that the decision is based upon more information about the student than that contained in the last response. This can be done by making available a response history of two, or more, previous responses. To do this a CAI system must be capable of keeping a cumulative record of responses in a form that makes the pertinent set available every time a contingency rule is to be used.

In addition to the availability of response histories, the CAI system programmed for unfolding a learning environment also must store in readily available form the rules for sampling the frames and the medium (e.g., audio or visual) or media to be used. Another requirement is the ability to use some contingencies, but not all, with a particular student. Related to this is the requirement for the system to be able to change the contingencies for each student if his performance is not meeting expected standards. This is called the feed forward concept in control theory. The system looks ahead to expected results based on what has happened and what possibilities could eventuate. After looking at the future possibilities it selects the most desirable one and then proceeds to do those things necessary to bring it about.

One type of rule specifies the sample of frames. In a linear program the sampling rules are simple: all students take the same program and they all see the same set of frames in the same order. The sample rule is to use the entire set of frames.

In idiomorphic programming the learning environment is predetermined but not pre-specified. A predetermined learning environment can be generated in either one of the two different ways referred to in the previous paragraph or by using both ways. Each way uses conditional rules of the form, "If response 'A,' then use procedure α to determine the next display." The display that appears next is predetermined by a procedure which can generate not just one, but a number of different displays.

One way to implement this form of conditional is to use a procedure that varies the content (e.g., the numerals) but not the sequence (e.g., the algorithm). A second way to implement it is to use a procedure that varies the sequence, but not the content. A third way is to vary both. The first way makes the sequences or logic conditional; the third way makes both content and sequences conditional.

Another type of instructional strategy rule determines the time intervals. With a pre-specified learning environment of the linear type the strategy rule is simple since all students go from frame 1 to frame 2, to frame 3, etc., and in the same order. The rule is simple but the value can vary since the time intervals are determined by the student; he sets his own pace. By pre-specifying the conditions the only individualization achieved by a linear program is in the rate at which each student progresses through the material. By pre-specifying the outcome of each response in an intrinsic program, all the branching possibilities are fixed, and each response alternative takes a student to a specific display. The contingency that is pre-specified in this form of programming is that between the last response and the next display.

The learning environment can be predetermined in still other ways. For example, the content of the next frames may be different depending upon the correctness of the student's responses to other frames on the same topic, and their sequence may depend upon aptitude scores.

With the predetermined level of functioning, the on-line decisions made by a teaching system change one, or more, of the teaching rules, or the teaching strategy. Two students, each making the same response to frame 105, for example, might see different displays, experience different time intervals, and have information available for different lengths of time, etc. The information display, the sequence of displays in the set, and the timing conditions would be contingent upon specific information about the student's past performance. One question to be answered is what type of information can be used in making decisions. This is a research question to be answered ultimately by empirical data. Hypotheses can be formulated; for example, it is possible to include aptitude and personality test scores as well as sets of students' responses to instructional frames during learning in the "if" portion of the conditional statements. Whichever one, or combination of scores is used in specifying contingencies for instructional decisions will depend upon the educational programmer's hypothesis regarding the optimum strategy of instruction.

When a computer-based system provides a contingency-determined learning environment based upon individual differences in performance, it uses idiomorphic programming. In teaching a set of concepts the rules which determine the order in which they are presented may change, for example, from an inductive sequence to a deductive one. While any rule may be changed, it might be useful to consider examples in which only one is changed. For instance, only the rules determining the information sequence might be changed. Another example might be the rule specifying the evaluative feedback contingencies (given the correctness of the learner's responses). If the student's performance is not up to expectation in correctness or in response latency, the system may change the rules used to teach him. For example, the rule being used may have been to give the student reinforcement every time he responded within a certain period of time. If he did not respond quickly enough, the reinforcement may be given on only half of the times he responded. This type of teaching rule would specify the contingencies between response latency and the subsequent reinforcement schedule. If the level of the student's performance is below that expected, easier frames may be presented to him. This last example illustrates yet another type of rule, one making the difficulty of the subsequent material contingent upon the number of errors.

These examples illustrate the types of decisions that can be made when contingencies are predetermined but not pre-specified. Instruction is more individualized when instructional rules are contingent upon both the student's performance and what is expected of him. Whenever one or more aspects of the student's actual performance do not match expected performance, a new rule is used. When a new rule is used, sets of displays and/or frame sequences become available that would not be available if the system operated at the pre-specified level of a linear or an intrinsic program.

Bases for Contingencies in Instruction

This analysis of teaching as the control of the learning environment is based upon the idiographic contingency model of tutorial instruction (Stolurow, 1965b, 1965c). This model distinguishes two types of contingencies. At the lower level there are event contingencies; at the higher level there are rule contingencies. The model is made operational in SOCRATES in three ways.

Student data possibilities. The first is in the number of possibilities which the system provides for the use of data in making decisions concerning the content and strategy to be used in the instruction of individual students. For instance, the student's ability, aptitude, and personality, as well as his knowledge or skill, can be used in making decisions at any point in the program. Patterns of responses, that is, various patterns of error and time, made during learning are included in the possible student characteristics which the system can use. Sets of these patterns are possible bases for making the subsequent experiences contingent upon past performance. It is possible to use pretest performance along with error and time data on selected frames as the basis for content and/or sequential contingencies. Consider an example of how these data might be used to make different decisions for each of two students. One student may have made a high pretest score and then gone through the program rapidly, making lots of errors; another student may have made a low pretest score, but gone through the program in the same way. The first student might be assumed to be careless, and, therefore, given instruction designed to make him more careful. The second student could be assumed to have failed to learn because of insufficient ability and therefore, to be in need of supplementary or remedial instruction.

Different display possibilities. The earlier example in which the system provided different consequences to students who made the same response to a particular frame illustrates a second way a computer-based system can make instruction more individualized: namely, by providing different display possibilities for every pedagogically-significant contingency. The system's adaptivity is increased by including in its library a large number of possible displays for each response possibility. The number of possible displays is increased in different ways. One way is by writing sets of materials at different levels of difficulty, another is by using different forms of encoding to communicate information, e. g., verbal versus mathematical. A third way is to separate the informative material that is intended to explain the answer from the different kinds of evaluative feedback. The example of the student who was responding carelessly is a situation in which the availability of separate types of feedback would be useful. With the separate types available in this system, it would have been possible for the system to tell him he was careless while also telling the other student who made the same response that he was improving. Furthermore, each student could be given his subsequent instruction with another teaching strategy. For example, the careless one could be required to make correct responses before being allowed to go on. The other student in the example, however, could be branched to different evaluative feedback, possibly of an encouraging type, and then given easier material so he would have a higher probability of being correct.

Either an instructional theory or empirical data specifies each of the relationships between a set of student characteristics and a set of display possibilities; consequently, the more display possibilities in the memory of the system, the more contingencies it can handle and the more individualized the instruction could be.

The differences between the first and the second way in which adaptivity is achieved lie in the range of unique display possibilities which each contingency rule could generate. Furthermore, the actual number of display contingencies is larger than the number of frames since both the temporal interval between a response and a display and the duration of each display can be controlled separately. Additional contingencies are possible through the preparation of alternative materials at different levels of difficulty or of readability so that comprehension and speed can be maintained at each student's optimum level. For example, the careless student might be shown a display for a longer period of time before the system accepts his next response, or he may have to wait longer for the next frame after he has responded.

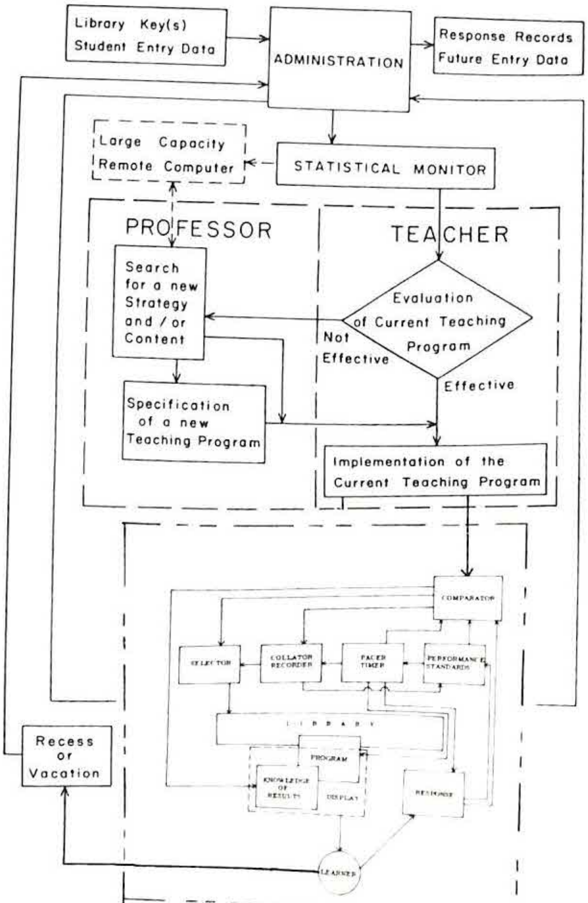
By programming the following conditions as contingencies, they can be used selectively in teaching: first, the difficulty or readability level of the information frames; second, their sequence; third, the temporal interval between responses and feedback as knowledge of results; fourth, the type of feedback statement used in terms of both their qualitative and evaluative nature; fifth, the schedule used to give feedback to the student; and sixth, the time interval between knowledge of results and the new program frame.

Criterion possibilities. A third way in which idiographic programming can make a system more adaptive is by using a variety of criteria in making decisions to either maintain or change the instructional rules themselves. A flow diagram describing a system for doing this is presented in Figure 3.

The program which allows the system to change rules on line is called the professor function. Basically this is the means for determining rule contingencies. It is to be contrasted with the ability of the system to function according to a fixed set of rules, which is called the teacher function. The teacher function is the part of the executive program that specifies the conditions for showing each one of a set of displays. The professor function is the part of the executive program that specifies the conditions for showing one set rather than another. These two functions are represented in Figure 3.

The professor function also requires ''if... then...'' rules, but these are rules that make the use of lower order rules contingent upon specific

Figure 3. System of criteria for decision-making in idiographic programming.



conditions. The professor rules imply teacher rules. In other words, if a student performs either better or worse than expected when a particular teaching strategy has been used, then some, or all, of these rules would be changed. This means that at least one other teaching strategy must be available as an alternative to the one being used. When an alternative is used, frames may be displayed that would not be displayed if the previous strategy had been used continuously.

An example of a pair of strategy rules which could be used by a system on an either/or basis would be correction procedure with an inductive sequence rule or with a deductive sequence rule. In using these strategies, it is necessary to have the conditions specified and a criterion available so that it is possible to determine if the student failed to meet the performance criterion. Assuming that he failed to meet the criterion when he was taught by the inductive strategy, he would be shifted to a deductive strategy in the next set of frames. In accomplishing this change, the criterion for the professor function might be ten consecutive correct responses none of which takes more than half a minute.

Whenever the student is being taught by one strategy rule, the teacher function is in control. However, the shift from one teaching rule to another is accomplished by the professor function which becomes effective whenever student responses fail to meet a criterion. Every time the student's performance fails to meet the professor criterion, one or more rules of instruction are changed. If a teacher criterion is not met, then the decision that is made is specified by a particular rule. The difference between the teacher and professor functions can be made more explicit by indicating that the use of one teaching rule may utilize a fixed set of instructional and feedback frames whereas the use of two or more rules may utilize additional informational and/or feedback frames. Therefore, with an executive program that can make rule changes and use different criteria for the changes, students may see materials when one rule is being used that they do not see if another one is being used.

Two types of criteria can be used in making decisions about instructional rules. One type of criterion is a predetermined value; it may be a value for errors, for time, or for a combination of both of these. It is illustrated by a conditional statement such as "If the student makes three or more correct responses to this set of frames, then continue with the use of rule x." The other type of criterion is generated by an algorithm, or computational formula, on line. This latter type of criterion is not predetermined; rather the algorithm and the nature of the data it uses are predetermined. Student response data are processed as they occur and the result of the processing is the criterion value used in deciding to continue or discontinue using a rule. The following is an example using response rate. If the student's rate of response increases so that the values fit a linear incremental function within a specified range of variation in slope, the strategy that is being used to teach him is maintained. If his rate of response does not fit a linear function within the range of permissible variation, the strategy shifts to the use of stronger positive evaluative feedback.

Another example of a computed criterion is the use of a mean value based upon the performance of students who have just been taught. As each student is taught, his data are used in the re-computation of the mean and standard deviation which become the new criterion in making a decision with the next student. In this sense the system has a memory, for it accumulates data on students it teaches and it adjusts its future performance based upon the way in which it is programmed to use the accumulated data. These are some of the ways in which SOCRATES can utilize response history and criteria to determine its future operation. Thus it functions as a cybernetic system. A complete record of each response and its latency is available. These data can be summed across topics or across students as desired to obtain a performance measure that is read out or used internally by the system for educational or research reasons. Furthermore, all records are kept for a student rather than for a station. In other words, a student can move from one station to another and the system keeps all his records together. He can stop and start up again at any time and the system has immediate access to all of his test scores and past responses that it was programmed to save.

IMPLEMENTATION OF THE MODEL

In order to accomplish the teaching functions of the idiographic contingency model, it is necessary to prepare instructional materials, including tests, and to write computer programs to manage the instructional materials in order to collect the data and to process it.

LOADING THE SYSTEM

The preparation of instructional materials for use in SOCRATES involves two processes. One is photography; the other is computer programming. The student looks at photographed images of printed or drawn materials. These can be relatively conventional programmed instruction frames, since pre-tutorial decisions can be based either upon extant standardized tests of achievement, ability, aptitude, and personality, or upon specially-prepared tests. A convenience of SOCRATES, which does not exist for some other computer-based systems, is that any test material prepared in conventional form can be photographed and used on the system. Processing materials to make them actually available under computer control can be accomplished relatively quickly since they simply need to be photographed and then a program can be written to control the duration and sequence of their display. The students' responses are recorded and scored automatically.

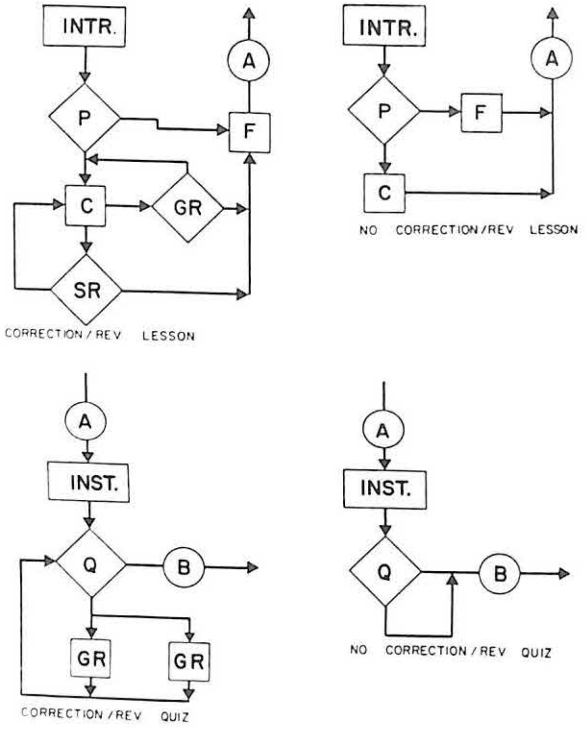
When materials are loaded on SOCRATES there are two parts to the content of its library. One is the displayable information stored on 35 mm. film for which there is a maximum reel capacity of 1,500 frames; the other is the program in the memory of the computer, both core and disk, in which the instructional strategies and student data are stored. It is necessary to write a computer program to control the presentation of each frame of film and to process any data to be used on-line. With computer-based systems that use a CRT display, however, it is necessary to write programs which generate the test items and the displays of program frames.

A more interesting use of SOCRATES occurs when educational programs are prepared especially for the system. It is possible to introduce a great deal more variation in the logic of instruction with the system than otherwise would be possible. Figure 4 presents a set of very simple flow diagrams associated with some research conducted with programs especially prepared for SOCRATES (Merrill and Stolurow, 1966). It also illustrates some of the discriminations that are made among materials to provide opportunities for flexibility in programming. P frames are those frames which represent programmed materials. F frames are feedback frames. C frames are frames which correct the student. GR frames are frames which provide for general review of the point covered in the P frame, and SR frames are frames which give specific reviews to the student. By providing these different types of frames in the film library, it is possible to route the student through selected subsets of them in whatever sequence seems most desirable. A computer program must be written to do this. It would be written to provide for the contingencies of a set of instructional rules.

The upper left panel of Figure 4 is a flow diagram in which a set of these frames is used to provide correction and review in a lesson. By correction procedure, we mean that the student must make a correct response to the material covered in the P frame before he is allowed to go on to the next P frame. If, however, it is desirable to route him through a review frame on this same material when the student responds in a way which the educational programmer feels indicates a review, the computer program is written to do this. The no-correction procedure is illustrated in the upper right panel. In this case the student is allowed to go on to the next frame whether he is right or wrong; however, he is first shown an F frame which tells him he is either correct or incorrect. The use of a no-correction rule is typical of a linear programming strategy.

The lower half of Figure 4 is a flow chart of the correction procedure and of the no-correction procedure as each is applied to questions in ordinary multiple-choice testing. The Q frame, or question, now replaces the P frame. The arrangement of P and Q frames on the film can be set up in whatever manner seems most efficient for use with the strategies that are intended for the students. The location of the frames in relation to each other on the film makes it possible to locate quickly the frames that could follow each P frame. The reason for this is to minimize the interval between the presentation of one frame and the retrieval of the next.

Figure 4. Program flow diagrams illustrating the experimental conditions of correction/review and no correction/review on lessons and quizzes.

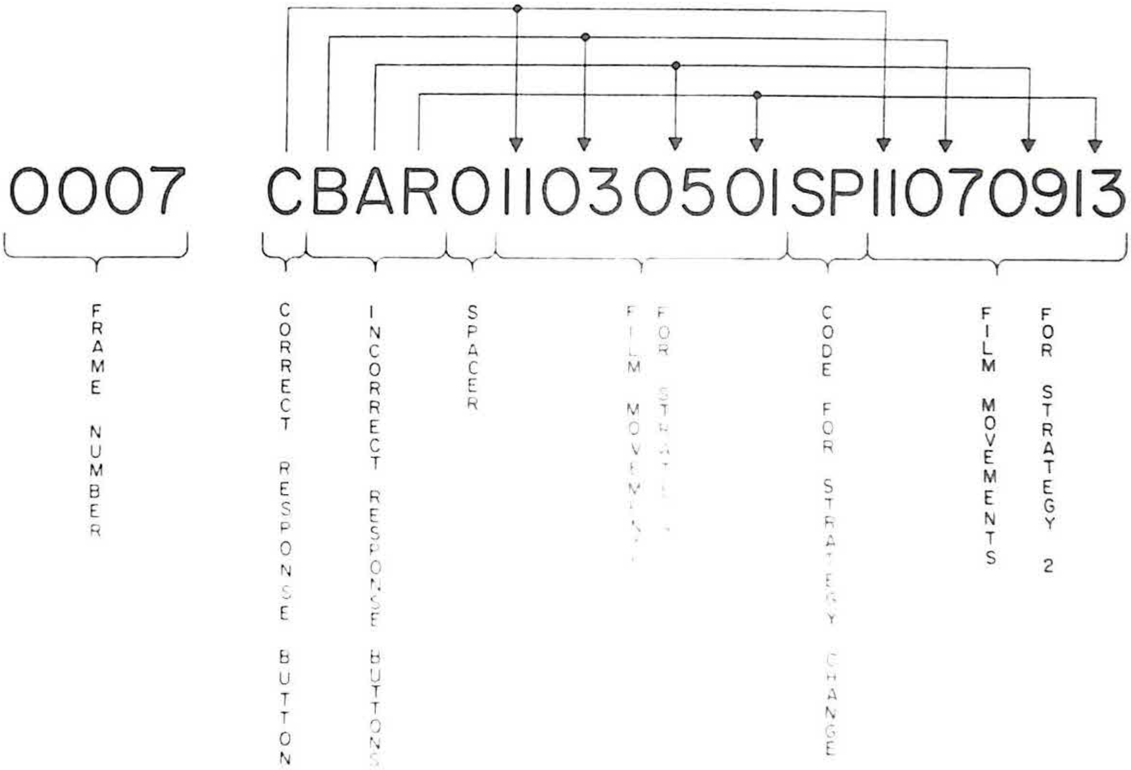


The decision rules to accomplish the various strategies are frequently presented initially in flow charts such as those in Figure 3. A group of one or more teaching rules may be required to implement a teaching strategy. Similarly, a set of decision rules or of teaching strategies comprises a teaching logic. The use of one or more strategies can be thought of as being employed at different points in the student's program to define the teaching strategy used with that program. If a particular set of contingency rules is used to provide the instructional strategy and a set of one or more strategies is used to make up the teaching logic, then the system is more or less adaptive to individual student needs depending upon the extent to which the particular set used to teach a student was determined before the student started. If the set was determined prior to instruction, the system is not as adaptive as if it were determined while he was learning and depended upon his performance during learning.

The research problem in making the system adaptive in an effective way is to determine the contingencies to use and the criteria so that the computer-based instructional system is programmed to implement those contingencies that instruct most efficiently. This is the primary problem in the further development of the idiographic contingency model. In order to solve this problem it will be necessary to begin by using the systems in a less adaptive level than they are capable of performing. Students with different characteristics will be assigned to strategies to determine their relative performance.

Code line of library table. A code line of a library table is presented in Figure 5. It represents one of the two sets of computer materials used to control instruction. The library table consists of lines of code which determine the frame locations and the various legal buttons which the student may

Figure 5. A code line from a library table.



use in responding to displayed material. The illustrated code line presented in Figure 5 is for instructional frame number 7 of a program. This is indicated on the left by the number 0007. Associated with each instructional frame is a set of lettered alternatives each of which is specified by a rule that is part of a teaching strategy. Each letter in the code line is associated with a frame address to which the student will be taken if he makes that response. If, for example, he makes the correct response, "C" in this example, the film is moved eleven framed forward. "C" is a correct response. The correct response alternatives are always listed before the incorrect ones, and after the first one the other correct alternatives are starred. There is only one correct response in the example. After "C," the letters "B" "A" "R" appear. Each is an allowable incorrect response. If any other key is used it results in no change in the display.

In addition to the library table there is the computer program which actually manages the movement of the program frames and the times of display. The various strategies and levels of adaptivity previously described can be accomplished through the use of these two sets of materials.

RESEARCH

Some of the research which relates to the idiographic model follows. While the model specifies the types of contingencies that appear to be sufficient for the design of adaptive learning environments, the exact nature of the contingencies which will maximize the rate of learning remains to be determined by research. Some studies which have provided information about contingency relationships can be described briefly. Since they are insufficient to document all of its facets, the studies mainly provide support for the general validity of the model.

THE EFFECTS OF SEQUENCE UPON APTITUDE REQUIREMENTS FOR LEARNING

A matrix of inter-correlations between a particular set of aptitude scores and the gain scores resulting from a learning experience indicates the extent to which the gains can be predicted from the various aptitude scores (DuBoise, 1962). The significance level of each correlation indicates whether the aptitude is or is not a significant factor in determining the gain score which students make. If students are randomly assigned to different teaching strategies, and a single set of aptitudes is related to the gains made, then differences in the correlations indicate that an aptitude is involved in learning under one strategy but not under another. This type of information relates to the diagnostic aspect of the model in that the correlations would identify the aptitudes that could be used in assigning students to the instructional strategy that will result in the greatest gains. If the strategies consist of rules that determine the frame sequence, then the students' aptitudes could be used to determine their optimum sequence.

One characteristic of teaching materials which can be made contingent upon aptitude is the sequence of frames presented to the student. If studies show that the size of the correlations between particular aptitudes and learning scores varies when different sequences of P frames are used so that some students do well when they learn with one sequence while others do well with a different sequence, it can be assumed that there is a rule that determines the sequence which would be best to use for a student who is relatively high in a particular aptitude and relatively low in another. Research revealing that aptitude-gain correlations do change when different strategies are used provides support for the idiographic contingency model. If aptitude-gain correlations do not change significantly, then this type of model would be inappropriate.

Asynchrony as a measure of sequence. One problem that arises in using a diagnostic-prescriptive approach is the lack of critical information. Critical information might be obtained through the use of variables in research that can be implemented in a selective way. For example, if frame sequence is to be manipulated selectively then it is important to have the rules for manipulating sequences so formulated that they can be implemented easily. Therefore, the way in which the sequence of frames, or displays, is characterized is important. The concept of *asynchrony* as applied to critical and noncritical stimulus elements appears to be quite useful for this purpose (Detambel and Stolurow, 1956). Asynchrony can be exemplified in the following illustration involving color, size, and shape. If all frames in a program use color, size, and shape of a set of objects and the student is learning a concept involving one of them, such as color, then color is the critical cue. Consider some of the things that happen in going from frame to frame in a sequence. The color may change from one frame to the next – for example, from red to blue – but the size and shape may not. In this example, not only are color and size asynchronous with respect to one another, but also color and shape are asynchronous. If this pattern of asynchrony existed in a program, then as the learner went from one frame to the next, color would change but neither size nor shape would change. It is highly unlikely that this set of relationships would continue indefinitely, but that is not the point being made at this time. The point being made is that for a series of frames it is possible to look at changes in critical characteristics and determine what other characteristics also change or do not change. In this way it is possible to determine the number of pairs of changes in stimulus characteristics which take place in going from one frame to the next one. If one characteristic, such as color, is critical to what is being taught, and other characteristics are not, then two different types of asynchronous sequences can be identified. One consists of asynchronous sequences in which a critical feature changes but a noncritical one does not. This can be called a CC type of asynchronous sequence since the critical characteristic changes. A second asynchronous sequence exists when a critical feature does not change but a noncritical feature does. This can be called a CNC type of asynchronous sequence.

Two opposite types of sequences also can exist. These are called *synchronous* sequences. One type occurs when both a critical feature and a noncritical feature change; the other occurs when neither changes, or, in other words, the display is repeated. These ways of characterizing sequences have been found to be useful in organizing materials in several studies (e.g., Detambel and Stolurow, 1956; Stolurow, 1956; Stolurow, Fogel, and Fogel, 1966; Anderson, 1966) and recently in Germany by Kotter (1965).

These concepts are treated more extensively in Stolurow (1956) and in Garner (1962).

The use of class-descriptive cues. The measures of asynchrony, while reliable for a particular frame sequence, do not, however, identify a unique sequence of frames. Using the same set of frames it is possible to make up different sequences which have the same amount of asynchrony. In other words if a sequencing rule were used by a computer to generate an asynchronous sequence the computer would not necessarily generate the same sequence each time the rule was used. However, in all cases the frequency with which CC and CNC types of sequences occurred would be the same. Since this is so, it is useful to differentiate among sequences with the same asynchrony value. A way of doing this is to use the concept of class-descriptive cue (Stolurow, 1956); any feature that does not change from one frame to the next frame within a subset of two or more frames of a program can function as a class-descriptive cue. Therefore wherever there is more than one feature that can be useful to the learner, the frames should be organized into subsets with a class-descriptive cue that is related to one of the objectives of instruction. The data thus far collected indicate that the use of class-descriptive cues in organizing displays does facilitate learning (Stolurow, 1956; Wulff and Stolurow, 1957).

The effects of sequence changes on the relationship between aptitudes and gains. An asynchrony rule in combination with a cue-organizing rule can produce sequences that provide class-descriptive cues for the learner. This teaching logic appears to be a useful way to synthesize a program of instruction.

Two studies will be reported here to indicate the results of this approach. In both studies analyses were made of the data to determine the resulting relationships between aptitude and gain scores when different sequences of the same frames were used for instruction.

The first study was conducted by Cartwright (1962). He used a fractions program which was presented to mentally-retarded students during secondary-level instruction. Two variations of the instructional material were prepared. In one program there was a systematic increase and decrease of the numerator and then the denominator of the fractions presented to the student as he went from frame to frame. In the other alternative, which was made up with the same frames, the sequence used did not maintain this relationship. The comparison of the two groups was therefore between a program with a

TABLE 1

CHANGES IN CORRELATIONS BETWEEN APTITUDE AND LEARNING SCORES FOR DIFFERENT SEQUENCES WITH EQUAL AMOUNT OF ASYNCHRONY (CARTWRIGHT, 1962)

Individual Characteristics	Sequence with Systematic Increment and Decrement (N = 20)	Sequence with Non-systematic Increment and Decrement (N = 16)
Mental Age - WISC		
Verbal	.02	-.26
Performance	.26	-.50*
Total	.17	-.49*
Mental Age - Binet	.18	-.49*
Language Aptitude	.12	-.54*

*p < .05

class-descriptive cue and one without a class-descriptive cue. Some of the data are reported in Table 1.

Cartwright found, for example, using a total intelligence test score from the WISC, that a .17 correlation existed with final achievement or gain scores under the class-descriptive cue (CDC) condition; whereas a correlation of -.49 was achieved with a sequence that did not have a class-descriptive cue. With the Binet intelligence test scores comparable correlations were obtained, namely, .18 and -.49. Even more interesting is the pair of correlations with the verbal aptitude test: a positive correlation of .12 was obtained with the CDC sequence and a -.54 with the other sequence. These data indicate that verbal aptitude becomes more or less significant as a contributor to learning depending upon the sequence in which the frames of the program are presented.

The difference in the correlations for the two sequences indicates that it is possible within a 30-point IQ range, from 53 to 83 in this case, to predict from either the general intelligence measure or the verbal aptitude measure what the residual scores of students will be under the non-CDC sequence. This means that with the CDC sequence neither intelligence nor language aptitude plays a significant role in determining the residual scores within this range of ability. With another sequence, however, these abilities played a more important part in determining what the students learned.

A second study, dealing with sequence differences and also employing a fractions program, but a different one, was conducted with normal students in elementary schools in Champaign, Illinois. This study was completed recently and revealed a pattern of relationships between aptitudes and learning as a result of sequence differences which was similar to that reported in the Cartwright study. Table 2 summarizes the set of correlations between pairs of variables which differed significantly for the two degrees of asynchrony. While there were other significant differences between the effects of these two sequences in addition to those shown in this table, these differences illustrate the point that changes in asynchrony can alter the aptitude requirements for achievement.

The data from both studies support the idiographic contingency model described earlier. They indicate that effective individualization of instruction can be accomplished by using teaching rules that offer instructional sequences which fit the aptitude characteristics of each student. The use of sequencing, and particularly of asychronous sequencing with CDCs needs to be studied further to determine the sequencing rules which are so related to the students' aptitudes that the best match for maximizing his performance is achieved. The research problem facing us now is to determine the specific kinds of relationships one needs to take into account in making the best match between aptitudes and rules for generating particular sequences of frames.

THE EFFECTS OF VARIATIONS IN ENCODING ON APTITUDE REQUIREMENTS

Another aspect of a teaching strategy is the form of symbol used to express a set of concepts or, in other words, the encoding employed to communicate the concepts to students. A recent study (Stolurow, Frase, and Odell, 1966) of encoding differences employed two different sets of notation in teaching logic to college students. One program used the Peano-Russell notation system, the other the Polish notation system. In all other respects the two programs were identical. Thus, any variations in outcome could be attributed to the encoding differences.

The first set of results (Table 3) indicates rather clearly that under the Peano-Russel notation system American students perform in a manner that is predictable from a set of aptitude tests. Comparable

TABLE 2

CORRELATIONS BETWEEN IQ AND PERFORMANCE AND BETWEEN SPECIFIC APTITUDES AND PERFORMANCE FOR DIFFERENT DEGREES OF ASYNCHRONY (UNPUBLISHED STUDY)

Variables	Sequence with Less Asynchrony	Sequence with Higher Asynchrony
California Test of Mental Maturity		
Total IQ and Gain Score	+.01	+.73*
Non-verbal IQ and Gain Score	-.13	+.55*
Iowa Test of Basic Skills		
Spelling and Gain Score	-.32	+.56**
Stanford Achievement Test		
Arithmetic and Gain Score	-.46	+.70**

* and **: The two correlations are significantly different from each other; *$p < .05$; **$p < .01$

TABLE 3

CORRELATION OF DIFFERENT APTITUDES WITH PERFORMANCE ON A TEST COVERING IDENTICAL CONTENT TAKEN AFTER LEARNING THE POLISH OR PEANO-RUSSELL NOTATION

Group	Vocabulary	Aptitudes for Mathematics	Inductive Reasoning	Inference
Peano-Russell	.27*	.31**	.31**	.01
Polish	.09	.18	-.01	.01

*$p < .05$ **$p < .01$

students, when trained with the Polish notation system, however, achieved gains which did not correlate significantly with the same aptitude tests. Thus, it appears that these aptitude measures are predicting the processing skills which are required of the learner by the symbol system used to encode the concepts and not the actual learning of the logical concepts themselves. This seems to be a rather clear finding since the concepts of logic were identical in the two programs. The processing skills which are required to decode the symbols representing these concepts appear to be quite different. Decoding the Polish expressions requires that a person trained to process English and mathematics statements be able to transpose symbols in order to make sense out of the displays presented to the students. These data indicate that one can teach the same set of concepts to students by two different, but presumably equivalent, encoding systems and in doing so can change not only the aptitude requirements for learning but also the transfer effects which the training produces. The data also indicate that the Peano-Russell notation system is superior to the Polish system for students trained in processing the English language.

THE EFFECTS OF VARIATIONS IN EVALUATIVE FEEDBACK ON PERSONALITY REQUIREMENTS

Still another set of contingencies pertaining to the model are those relating personality characteristics of the learner to his performance gains when different conditions of evaluative feedback, or social reinforcement, are employed. Social reinforcement in these studies means evaluative feedback statements which tell the student about the quality as well as the correctness of his response. Thus, approval could be expressed in evaluative statements such as, ''Excellent, you are coming along fine,'' or ''Right on the button''; for negative evaluation, or reproof, evaluative statements such as ''No, you are not thinking,'' ''No, what a silly mistake to make'' express the attitudes.

Each of these kinds of statements was given to different groups of students all of whom studied the

TABLE 4

CORRELATIONS BETWEEN PERSONALITY VARIABLES (a) PERFORMANCE, (b) ATTITUDE AND (c) TIME UNDER FOUR DIFFERENT CONDITIONS OF SOCIAL REINFORCEMENT (FRASE, 1965)

Group	Aggression/ Performance	Aggression/ Attitude	Deference/ Attitude	Deference/ Time
Condition 1 ±	.15	.24	-.05	.07
Condition 2 +	-.44*	-.50*	-.06	.48*
Condition 3 -	.35	.73***	-.59*	.37
Condition 4 0	-.50*	-.16	.00	-.40

$*p < .05$ $***p < .005$

same basic program in logic (Frase, 1963). Group one (+) had correct and incorrect responses evaluated. Group two (+) had only their correct responses evaluated. Group three (-) had only their incorrect responses evaluated. Group four (0) had no evaluation of either one, just knowledge of results.

In Table 4, three sets of correlations are of interest. One set is between personality test scores and learning scores (e.g., aggression/performance); a second is between personality scores and attitude (e.g., aggression/attitude); and a third set is between personality and learning time. All three sets revealed that each of the four experimental conditions of social reinforcement was most suited to a different personality group. Three personality characteristics, aggression, deference, and intraception (Murray, 1938) were found to be important in relation to performance under the four different conditions of social reinforcement. For example. students who were high in aggression performed best when they were given negative evaluation of their wrong responses and not given positive evaluation of their right responses. Under two conditions – number 2, positive evaluation, and number 4, no social reinforcement – aggression was related negatively to performance. In other words, for the students high in aggression, performance was impaired under these conditions. Under condition number 3, negative reinforcement, aggression was related positively to attitude but unrelated to achievement. But the relationship with performance, although positive, was not significantly different from zero.

Under the punitive conditions (number 3) high deference was related to lower attitude but was not related to achievement performance.

Intraception was negatively correlated with speed and efficiency, and related to high attitude under maximum social reinforcement, condition number 1.

In another study by Parisi (1965) the basic paradigm of the Frase study was replicated and an additional personality test, the Cattell 16 PF, was given prior to learning. All materials were administered in Italian to Italian students in Italy. These results support those of Frase that negative evaluations were more important that positive evaluations. Furthermore, positive reinforcement was more effective than negative reinforcement in reducing the correlation between intelligence and performance. Performance on the achievement test was related to deference, or lack of need for autonomy, when social reinforcement was used. When no social reinforcement was used, performance was related to need for achievement and exhibitionism.

The data from these studies seem to show that it is possible to alter the quality and speed of the performance as well as the attitude of students by the way in which social reinforcement is used during the learning experience. They further suggest that it might be possible to achieve an optimum matching of personality characteristics with social reinforcement conditions for each individual student so that the general level of effectiveness of instruction can be raised by individualizing these contingencies.

SUMMARY

This paper summarizes a general model used to design a computer-based instructional system. This idiographic contingency model defines classes of variables that are presumed to be important in adaptive instruction. It also specifies relationships between those variables that characterize the learner before and after learning which can be used selectively to individualize instruction. Included among the "before" measures are knowledge, aptitude, and personality. The data indicate that the "before" measures have different relationships with the "after" measures (e.g., amount learned and attitude) under different conditions of learning. These data support the model.

The kinds of decisions that need to be made to individualize instruction were discussed and a set of research studies relating to them was described briefly. Relevance of the data to the model was indicated and it was pointed out that learning environments can be made more adaptive by using appropriate matching of: (a) aptitude and sequence contingencies; (b) aptitude and encoding contingencies;

TABLE 5

CORRELATIONS BETWEEN INTELLIGENCE SCORES AND PERFORMANCE VARIABLES – (a) MIDPROGRAM TEST, (b) POSTTEST AND (c) ERRORS IN PROGRAM – UNDER FOUR DIFFERENT CONDITIONS OF SOCIAL REINFORCEMENT (PARISI, 1965)

	16 P. F. Factor B[b] Intelligence with			Verbal Reasoning with		
Group[a]	LM	LP	E	LM	LP	E
Condition 1 ±	-.07	.33	-.22	-.15	.05	-.24
Condition 2 +	-.16	.51*	-.17	.40	.37	-.31
Condition 3 -	.38	.49*	-.44	.40	.49**	-.23
Condition 4 o	.48*	.42*	-.46*	.47*	.46*	-.18

[a] The social reinforcement conditions are coded as follows: Condition 1± mean positive and negative social reinforcement; Condition 2+ is positive only; Condition 3- is negative only; and Condition 4 o is neither positive nor negative.

[b] LM is logic midprogram; LP is logic posttest; and E is errors.

*$p < .05$ **$p < .01$

and (c) personality and evaluative feedback contingencies.

The technological capability for accomplishing a high degree of individualization in instruction exists. The problem now is to find the critical psychological and educational data which would make such a technological capability an effective means of instruction.

FOOTNOTE

1. This paper appeared as Technical Report No. 12 from the Training Research Laboratory, Office of Naval Research, Contract NONR 3985 (04). Urbana, Illinois: University of Illinois, 1966.

REFERENCES

Anderson, R. C. Two rules for manipulating the sequence of stimuli in concept learning tasks. Journal of Educational Psychology, 1966.

Atkinson, R. C. Instruction in initial reading under computer control: the Stanford project. Journal of Educational Data Processing, 1967, 4 (4), 175-192.

Bushnell, D. D. Computer-based teaching machines. Journal of Educational Research, 1962, 55, 528-531.

Cartwright, G. P. Two types of programmed instruction for mentally retarded adolescents. Unpublished master's thesis, University of Illinois, 1965.

Coulson, J. E. (Editor) Programmed learning and computer-based instruction. New York: Wiley, 1962.

Crowder, N. A. Automatic tutoring by intrinsic programming. In A. A. Lumsdaine and R. Glaser (Editors), Teaching machines and programmed learning: a source book. Washington, D. C.: National Education Association, 1960. Pp. 286-298.

Detambel, M. H. and Stolurow, L. M. Stimulus sequence and concept learning. Journal of Experimental Psychology, 1956, 51, 34-40.

Duboise, P. and Glaser, R. (Editors), Training research and education. Pittsburgh: University of Pittsburgh Press, 1962.

Frase, L. T. The effect of social reinforcers in a programmed learning task. Technical Report No. 11, Training Research Laboratory, University of Illinois, 1963, Contract NONR 1834 (36), Office of Naval Research.

Garner, W. R. Uncertainty and structure as psychological concepts. New York: Wiley, 1962.

Gentile, J. R. The first generation of computer-assisted instructional systems: an evaluation review. University Park, Pennsylvania: Computer Assisted Instruction Laboratory, Pennsylvania State University, 1965.

Hansen, D. N. Computer assistance with the educational process. Review of Educational Research, 1966, 36, 588-603.

Kötter, L. Der Einfluss unterschiedlicher Signalfolgen aud den Konzepterwerb (minimale und maximale Variation der Signale). Bericht Über den 24. Kongress der Deutschen Gesellschaft fur Psychologie. Göttingen: Verlag fur Psychologie, Dr. C. J. Hogrefe, 1965.

Lippert, H. T. Computer aided learning and transfer effects of Russian pronunciation. Unpublished doctoral dissertation, University of Illinois, 1967.

Markle, S. M. Good frames and bad: a grammar of frame writing. New York: Wiley, 1964.

Merrill, M. D., and Stolurow, L. M. Hierarchical preview versus problem oriented review in learning an imaginary science. American Educational Research Journal, 1966, 3, 251-261.

Murray, H. A. Explorations in personality. New York: Oxford, 1938.

Parisi, D. Social reinforcement and performance in programmed learning in Italy. Technical Report No. 27, Training Research Laboratory, University of Illinois, Contract NONR 1834 (36), Office of Naval Research, 1965.

Silberman, H., and Coulson, J. E. Automated teaching. In H. Borks (Editor), Applications in the behavioral sciences. Englewood Cliffs, New Jersey: Prentice Hall, 1962. Pp. 308-335.

Skinner, B. F. The science of learning and the art of teaching. Harvard Educational Review, 1954, 24, 86-97.

Smallwood, R. D. A decision structure for teaching machines. Cambridge, Massachusetts: M. I. T. Press, 1962.

Stolurow, L. M. Utilization of class-descriptive cues in the learning of technical information – studies in task engineering. In G. Finch and F. Cameron (Editors), Symposium on Air Force human engineering, personnel and training research, Technical Report 56-8. Baltimore: Air Research and Development Command, 1956. Pp. 248-266.

Stolurow, L. M. Teaching by machine. Cooperative Research Monograph No. 6, OE-34010. Washington, D. C.: U. S. Office of Education, 1961.

Stolurow, L. M. A model and cybernetic system for research on the teaching-learning process. Programmed Learning, 1965, 2, 138-156. (a)

Stolurow, L. M. Idiographic programming. National Society for Programmed Instruction Journal, 1965, 4, 10, 12. (b)

Stolurow, L. M. Untersuchungen zur kybernetischen Modellierung von Lehr und Lernprogrammen. Zietschrift für Psychologie, 1965, 171 (suppl.), 247-274. (c).

Stolurow, L. M. The computer-assisted instructional system in theory and research. In D. Unwin and J. Leedham (Editors), Aspects of educational technology. London: Methuen, 1966. Pp. 257-273.

Stolurow, L. M. Computer-based instruction: psychological aspects and systems conception of instruction. Journal of Educational Data Processing, 1967, 4(4), 193-215.

Stolurow, L. M., and Davis, D. Teaching machines and computer-based systems. In R. Glaser (Editor), Teaching machines and programmed learning, II: Data and directions. Washington, D. C.: National Education Association, 1965. Pp. 162-212.

Stolurow, L. M., Fogle, R. E., and Fogle, C. A. A comparison of two degrees of asynchrony. Technical Report No. 10, 1966, Training Research Laboratory, University of Illinois, NDEA Title VII Contract 4-20-002, U. S. Office of Education.

Stolurow, L. M., Frase, L. T., and Odell, S. J. Relative transfer effects of learning systems of notation for encoding expressions in a course of programmed symbolic logic. Unpublished manuscript, Training Research Laboratory, University of Illinois, 1966.

Stolurow, L. M., and Lippert, H. T. Prompting, confirmation and overlearning in the automated teaching of a sight vocabulary. In J. P. DeCecco (Editor), Educational technology. New York: Holt, Rinehart and Winston, 1964. Pp. 187-197.

Taber, J. I., Glaser, R., and Schaefer, H. H. Learning and programmed instruction. Reading, Massachusetts: Addison-Wesley, 1965. Pp. 65-72.

Wulff, J. J., and Stolurow, L. M. The role of class-descriptive cues in paired-associates learning. Journal of Experimental Psychology, 1957, 53, 199-206.

Zinn, R. Computer assistance for instruction. Automated Education Letter, 1965, 1, 4-14.

PART IV

Instructional Methods and Teacher Behaviors

TEACHING METHODS, or the system of interactions which occur between a teacher and students, have been the subject of many investigations in the last decades. In this Part, two recent paradigms for research on teaching are presented. N. L. Gage points out that the global research on teaching methods has not produced consistent, replicable or useable results. Therefore, a different and more limited approach must be used. In "An Analytical Approach to Research on Instructional Methods" he discusses micro-teaching, a technique which permits limited subsets of a comprehensive teaching system to be explored analytically. Micro-teaching permits the gathering of data on a limited set of student-teacher interactions during a short span of time, less than a class period. From such research "micro criteria" of effectiveness may be hypothesized and tested in terms of technical skills observable within the micro-teaching situation.

Another analytical approach to research on teaching is proposed in the paradigm set forth by Ned A. Flanders in "Interaction Analysis and Inservice Training." Flanders proposes a system of research in which a limited number of categories of student-teacher interaction is used in observing classroom interactions. Interaction data are recorded at 3 second intervals. The resultant matrix of data allows rather effective analysis of the verbal interchange between the teacher and the pupils. Interaction analysis is discussed in terms of its use as a research tool and as a means of effecting behavioral changes in teachers, such as inservice programs.

An Analytical Approach to Research on Instructional Methods

Nathaniel L. Gage
Stanford University
Palo Alto, California

INSTRUCTIONAL method constitutes one of the most important and promising but also the most frustrating of the areas of educational research and development. Compared with the areas of learning, subject matter, instructional materials, and organization for instruction, instructional method appeals to the classroom teacher as closest to the heart of her problem. It is all right for a teacher to know about learning, to know his subject matter, to have appropriate instructional materials, and to fit into a given organization for instruction. But what a teacher really wants to know is, ''What should I do in the classroom?'' If you ask prospective teachers or teachers on the job, ''Where do you really want help?'' I think the reply will deal with some aspect of instructional methods.

Unfortunately, the urgency of this demand has not been accompanied by corresponding success in meeting the demand. Research on teaching has yielded relatively few solid and usable results. The stature of theory and research in other areas puts them well ahead of the study of teaching in the struggle for scientific maturity.

Implicit in what I am saying is a basic distinction between research on learning and research on teaching. The former deals with all the conditions under which learning, or a change in behavior due to experience, takes place. And as I have already indicated, the study of learning is relatively mature, well established, with many volumes of substantial literature to its credit. Research on teaching, on the other hand, deals with a subset of the conditions under which learning occurs in one person, namely, the conditions established by the behaviors of another person, called the teacher. As our schools have developed during the past two or three millennia, we have always attempted to promote and improve the learning process through the intermediation of such another person. Until very recently, the assumption that teachers were helpful or even necessary for many important kinds of learning that society wanted to promote went unchallenged. Even today, the challenge of independent study or computer-assisted instruction and other devices is a mere whisper against the thunder of the assumption that teachers are necessary, that teachers are here to stay. My discussion of instructional methods is going to be based on that assumption. It is the relatively neglected, undernourished, and underachieving subset of research on learning which I call research on teaching that I shall deal with here.

RESEARCH ON TEACHING

The Stanford Center for Research and Development in Teaching is devoted to this problem area. In its conceptual framework, teaching, or the behaviors and characteristics of teachers, stands at the center. This domain contains variables that serve as both independent and dependent variables in the Center's program. When the teachers' behaviors and characteristics serve as independent variables, and the pupils' learning and behaviors serve as the dependent variables, then we have research on teacher effectiveness, or, more neutrally, research on teacher effects. On the other hand, when teachers' behaviors and characteristics serve as the dependent variables, and teacher education programs and procedures serve as the independent variables, then we have research on teacher education. Taking both research on teacher education and research on teacher effects as our domain, we have research on teaching, as it is understood in the Stanford Center. My subject is the Center's program of research and development on instructional methods, and my procedure will be to work from the past to the present, from the general to the specific, and then to try for a look at where we are going. When I get down to specifics, I shall be dealing primarily with research in which I am involved, rather than all of the research underway in our Center.

PAST RESEARCH

But first, let us look at where research on teaching has been. As the behavioral sciences go, it has a respectably long history but a regrettably inglorious one. Research on teaching has been going on almost as long as research on learning. Some studies were made in the 1910's and 1920's, and quite a few were made during the 1930's. By the early 1950's, substantial reviews and bibliographies of research on teaching began to appear. And during the last decade, the flow of research on teaching has indeed become significant. But the early years did not pay off in solid, replicable, meaningful results that had considerable theoretical or practical value. Positive and significant results were seldom forthcoming, and they survived replication even less often. The research yielded many findings that did not make sense, that did not hang together in any meaningful way.

Under these conditions, as Kuhn (1962) has pointed out, research workers are impelled to reexamine their first principles, the paradigms by which they guide their efforts. The model problems and solutions of the community of researchers on teaching were accordingly subjected to more and more reappraisal. Licking the wounds inflicted by their negative results, researchers on teaching built up a modest literature of new conceptual frameworks, approaches, and paradigms for research on teaching.

To illustrate, let me refer to one of the dominant paradigms that even today leads many discussions and research projects into the wilderness. This is the paradigm that says that what we need above all, before we can select and train better teachers, is the criterion of teacher effectiveness. Here is one example of that kind of approach:

> The lack of an adequate, concrete, objective, universal criterion for teaching ability is thus the primary source of trouble for all who would measure teaching. One typical method of attack used in rating scales is to compile a list of broad general traits supposedly desirable for teachers, with respect to which the rater passes judgement on each teacher. This amounts to an arbitrary definition of good teaching, which is subjective and usually vague, but it does not necessarily lead to an identification of it. Only if the traits themselves can be reliably identified can their possessor be identified as a "good teacher" according to the definition laid down in the scale. Even when the scale is made quite specific, relating not to general traits but to concrete procedure, the fundamental difficulty remains, that there is no external and generally accepted criterion against which the scale can be validated to establish the significance of its items (Walker, 1935, pp. x-xi).

This kind of writing implies that there is some magic variable that applies to all of teaching, for all pupils, at all grade levels, in all subject matters, and in all objectives. The phrase "the criterion of teacher effectiveness" betokens a degree of generality that has seldom been found in any branch of the behavioral sciences. It also reflects the mistaken notion that such a criterion, largely a matter of values, can be established on the basis of scientific method alone.

RECENT RESEARCH

The so-called criterion problem misled a whole generation of researchers on teaching, embroiled them in endless and fruitless controversy, and lured them into hopelessly ambitious attempts to predict teacher effectiveness over vast arrays and spans of outcomes, teacher behaviors, time intervals, and pupil characteristics, all on the basis of predictive variables that had only the most tenuous theoretical justification in the first place. It is little wonder that, when Berelson and Steiner (1964) dealt with the subject of teachers' behaviors and characteristics in their inventory of scientific findings in the behavioral sciences, they dismissed the "large number of studies" with the single dismal sentence that "there are no clear conclusions" (p. 441).

If the global criterion approach had proved to be sterile, what was the alternative? The answer was to take the same path that more mature sciences had already followed: If variables at one level of phenomena do not exhibit lawfulness, break them down. Chemistry, physics, and biology had, in a sense, made progress through making finer and finer analyses of the phenomena and events they dealt with. Perhaps research on teaching would reach firm ground if it followed the same route.

Apparently, a number of students of the problem had this general idea at about the same time. In 1962, writing my chapter on paradigms for the Handbook of Research on Teaching, I coined the term, "micro-criteria" of effectiveness. As I said in that chapter:

> ... One solution within the "criterion-of-effectiveness" approach may be the development of the notion of "micro-effectiveness." Rather than seek criteria for the over-all effectiveness of teachers in the many, varied facets of their roles, we may have better success with criteria of effectiveness in small, specifically defined aspects of the role... a sufficient number of laws applying to relatively pure aspects of the teacher's role, if such laws could be developed, might eventually be combined, ... to account for the actual behavior and effectiveness of teachers with pupils under genuine classroom conditions" (Gage, 1963, p. 120).

A group of workers at Stanford University, to which I was to move a few months later, took a similar view. In the Stanford program for training secondary-school teachers, Robert Bush, Dwight Allen, and their co-workers adopted what is now known as the technical skills approach. Technical skills are specific instructional techniques and procedures that a teacher may use in the classroom. They represent an analysis of the teaching process into relatively discrete components that can be used in different combinations in the continuous flow of the teacher's performance. The specific set of technical skills adopted in the teacher-education program at Stanford may be quite arguable. Indeed, the list of skills has been revised a number of times over the past few years. What is important is the approach – the attempt to analyze teaching into limited, well-defined, components that can be taught, practiced, evaluated, predicted, controlled, and understood in a way that has proven to be altogether impossible for teaching viewed in the larger chunks that occur over a period of an hour, a day, a week, or a year.

When analyzed-teaching, in the form of technical skills, is made the focus of our concern, we find it possible to do fairly satisfying research both on teacher education and on teacher effects. The satisfaction comes from being able to measure or manipulate relevant independent variables, perform true experiments, or make careful analyses and measure relevant dependent variables.

The idea of technical skills may be illustrated by the terms used in a recent list of such skills. One was called "establishing set," or the establishment of cognitive rapport between pupils and teacher to obtain immediate involvement in the lesson; one technique for inducing a positive set is the use of relevant analogies. A second technical skill is that called "establishing appropriate frames of reference," or points of view. A third technical skill is that of "achieving closure," or pulling together major points, linking old and new knowledge, at appropriate points within a teaching episode as well as at the end. A fourth technical skill is that of "using questions" in such a way as to elicit the kinds of thought-processes and behaviors desired such as simple recall, or concept formation, or evaluation. Other technical skills are those in "recognizing and obtaining attending behavior," "control of participation," "providing feedback," "employing rewards and punishments," and "setting a model."

MICRO-TEACHING

These technical skills into which important aspects of the teaching job have been analyzed are not merely the subjects of lectures and discussions in the teacher education program. Rather, they form the basis for the intern's practice teaching prior to his entrance into actual classrooms. This procedure, well known by now as "micro-teaching," consists in getting the trainee to teach a scaled-down teaching exercise. It is scaled down in terms of time because it lasts only 5 to 10 minutes. It is scaled down in terms of class size, because he teaches a group of not more than five pupils, who are brought in and paid to serve as pupils in the micro-teaching clinic. It is scaled down in terms of the task, since the trainee attempts to perform only one of the technical skills in any single micro-teaching session. The sessions are recorded on video tape, and the trainee gets to see and hear himself immediately after the session. While he looks at and listens to himself, he receives criticisms and suggestions from supervisors trained to be both perceptive and tactful. Then he "re-teaches" the same lesson to a new small group of pupils in an attempt to improve on his first performance of the specific technical skill that is his concern in that session.

Obviously, the general idea is subject to many variations. The size of the class can be manipulated; the number of trainees teaching a given group of children can be increased; the duration of the lessons can be lengthened; and the nature of the teaching task can be made more complex so as to embrace a group of technical skills in their real-life combinations. But the idea of analyzing teaching into technical skills remains the heart of the method and provides its power as a paradigm for research.

The research on micro-teaching and technical skills in the Stanford teacher education program has taken the form of experiments in which various procedures for feedback to the trainee are manipulated. Professors Dwight Allen and Frederick McDonald have organized a program of research on variables hypothesized to influence the learning of the technical skills of teaching. Their independent variables fall into three categories: practice variables, feedback variables, and demonstration variables. A practice variable may consist in micro-teaching versus teaching in an actual classroom. A feedback variable may be the positive or negative character of the feedback, or the mediation of the feedback by another person rather than the trainee himself. Finally, a demonstration variable may take the form of symbolic demonstration, consisting of written or spoken words, or perceptual demonstration, consisting in either live or video-taped portrayals of the desired behavior; and each of these can consist either of self-modeling or modeling by others. Other independent variables have been identified, such as the timing of a reinforcement, the amount of practice, and the amount of feedback.

This condensed description of the Allen-McDonald research program can suffice to illustrate the use of the analytic approach to research on teacher education. Their research takes the form of true experiments in which subjects are randomly assigned to different values of the independent variable.

TECHNICAL SKILLS APPROACH – EXPLAINING ABILITY

I should like to turn now to an example of the way in which the technical skills approach can be applied to the study of teacher effects. This research has dealt with a technical skill that I call "explaining," or the skill of engendering comprehension – usually orally, verbally, and extemporaneously – of some process, concept, or generalization. Explaining occurs in all grade levels and subject matters, whether it is a fifth-grade teacher explaining why the time in New York differs from that in San Francisco or a geologist explaining how the ice age may have been caused by volcanic eruptions. Everyday observation tells us that some people explain aptly, getting to the heart of the matter with just the right terminology, examples, and organization of ideas. Other explainers, on the contrary, get us and themselves all mixed up, use terms beyond our level of comprehension, draw inept analogies, and even employ concepts and principles that cannot be understood without an understanding of the very thing that is being explained. To some of us, it has seemed that explaining comes very close to being the inner essence of instruction, so that when a teacher is attempting to explain proportionality to his geometry class or irony to his English class, he is behaving more purely as a teacher than when he is attempting, say, to motivate, promote discussion, or maintain discipline. At any rate, we have made some studies of explaining ability in the attempt to determine some of the characteristics of effective explanations.

EXPLAINING ABILITY STUDY IN MICRO-TEACHING CLINIC

The first study was made in the micro-teaching clinic at Stanford during the summer of 1965 by J. C. Fortune, R. F. Shutes, and N. L. Gage (1966). We attempted to determine the generality of explaining ability, that is, the degree to which the ability to explain one topic was correlated with the ability to explain another topic, and the degree to which the ability to explain a topic to one group of pupils on one occasion was correlated with the ability to explain the same topic to another group of pupils on another occasion. We also were able to design the study so as to determine the degree to which there was generality over both pupils and topics, or the degree to which the ability to explain one topic to one group of pupils on one day correlated with the ability to explain another topic to another group of pupils on another day. Because there were only sixty pupils to be shared in groups of five among approximately forty interns in the micro-teaching clinic, the design became quite complex in order to avoid having any intern teach the same topic to the same group of pupils more than once and to avoid having the same group of pupils receive an explanation of the same topic more than once. Accordingly, the forty social studies interns – and we chose to work with the social studies interns only because there were more of them than any other kind of intern – were divided into five clusters of eight interns each. The lectures dealt with twenty different topics, each consisting of an "Atlantic Report" from the Atlantic Monthly. The correlations that we obtained were thus medians of five correlations, each based on four, six, six, six, and eight interns, respectively.

The index of lecture effectiveness, or what I would like to call the micro-criterion of teacher explaining ability, was the pupils' mean score on a 10-item test of their comprehension of the main ideas of the lecture, which was presented by each intern in 15 minutes under somewhat standardized conditions. This mean score was adjusted for the mean ability of the pupils in the given group as measured by their scores on all of the other topics. Similarly, any given mean score was adjusted for the difficulty of the topic as measured by the mean score of all groups of pupils on that topic. Thus, the variance of the adjusted mean posttest comprehension scores was attributable not to the ability of the pupils or the difficulty of the topic but rather to the differences among the teachers. We then investigated the question of the various kinds of generality by determining the median intercorrelations among the various means. The upshot of this part of the study was that generality over topics was nonexistent, and generality over groups was about .4. In other words, the interns were moderately consistent in their ability to explain the same topic to different groups on different occasions, but they were not consistent in their ability to explain different topics.

The study dealt with the correlations between explaining effectiveness and the pupils' rating of various aspects of the explanations. The pupils rated the interns' performance with respect to twelve items, such as clarity of aims, organization of the lesson, selection of material, and clarity of presentation. To us it seemed that some of these dimensions should correlate more highly with explaining ability than others. In particular, we hoped that such discriminant validity would be manifest in the form of a higher correlation between the mean rating of the lecture for "clarity of presentation" than for any of the other items of the Stanford Teacher Competence Appraisal Guide. Our hope was borne out; the correlation of the adjusted mean posttest comprehension scores with pupils' ratings of "clarity of presentation" was .56, higher than that with any of the other rating scale items. This result seems to us to support the validity of both the index of lecture effectiveness and the mean ratings by the pupils.

EXPLAINING ABILITY STUDY IN PUBLIC SCHOOLS

During the school year 1965-1966, I was able, in collaboration with Barak Rosenshine, to undertake a replication and extension of this study in the public schools (Rosenshine, 1968). Because there was no lack of students in the high-school classes, taught by their own teachers, we did not become involved in the complexities of design necessary in the micro-teaching clinic. To put it very briefly, we got forty eleventh-grade social studies teachers each to deliver a 15-minute lecture on an "Atlantic Report" on Yugoslavia taken from the Atlantic Monthly. The teachers had been given the article several days in advance, and had been told to prepare a lecture that would enable their pupils to answer a 10-item multiple-choice test of comprehension of the article's main ideas. To guide them in preparing their lecture, they were given five of the multiple-choice questions that would be asked, while the other five questions were withheld. After the 15-minute lecture, in which the teachers were permitted to use the blackboard but no other aids, their students took the 10-item test. They also rated the teacher's lecture on items similar to those I have already described. The next day, the same teachers and classes did the same things, except that the subject matter was an "Atlantic Report" on Thailand; again the teachers had been given five of the ten items as a guide to the kind of lecture that they should prepare and had been told to focus on the explanation of the major ideas, concepts, and principles brought out in the article, which constituted the curriculum for this bit of teaching. On the third day, the classes heard a third lecture, one that was the same for all classes, a tape recorded 15-minute lecture on Israel, a verbatim reading of an "Atlantic Report," and then the pupils again took a 10-item test based on that article.

The class mean on the Israel test was used to adjust the class means on Yugoslavia and Thailand for between-class differences in ability. Our reasoning was that the score on such a test of comprehension of a uniform lecture would be more useful in controlling relevant kinds of ability than would the usual scholastic aptitude test. The class means on Yugoslavia and Thailand were also adjusted for teacher differences in the content-relevance of the lecture, as determined by scoring the transcript of the lecture for relevance to the ten items on the comprehension test. We then assumed that the variance that still remained in the adjusted comprehension test means of the classes would reflect differences between the teachers in what we were concerned with, namely, the intellectual style and process of the teacher's lecture. In this study, the teacher's adjusted effectiveness index on Yugoslavia correlated .47 with his effectiveness on Thailand; i. e., there was considerable generality of effectiveness over topics, even after student ability and content relevance had been partialed out.

It should be noted that we were using the micro-teaching idea in this investigation. The teaching was restricted to just one aspect of the teacher's role, namely, ability to explain the current, social, political, and economic situation in another country. The curriculum was also scaled down. We also used another major feature of the micro-teaching clinic, the video tape recorders which made it possible for us to study the teacher's behavior, verbal and nonverbal, at leisure.

One major question was that of whether our criterion, or micro-criterion, of teacher effectiveness in explaining, namely, the mean comprehension score of the pupils, adjusted for both mean pupil ability and content-relevance, contained variance that would be manifested in something about the lecture that was visible or audible. In other words, was there some difference between good and poor explanations that was worth trying to analyze? So we picked two lectures on Yugoslavia that were extremely high on our index of effectiveness and two that were extremely low. We had a group of eight judges read the article on Yugoslavia and take the comprehension test, and then watch and listen to all four of these lectures. Then the judges ranked the lectures in terms of perceived effectiveness in engendering comprehension as measured by the 10-item test. It turned out that the judges' post-dictions were quite significantly more accurate than could have been expected on the basis of chance, and we were accordingly reassured that our micro-criterion was indeed reflected in something that could be seen or heard in the lecture.

But the major concern of this investigation was to determine the cognitive and stylistic correlates of the lecture's effectiveness. For this purpose,we used extreme groups to minimize the labor of scoring a host of variables about which we had no great conviction. So the ten most effective explanations on Yugoslavia were identified and also the ten least effective. From these, we chose at random five of the most effective and five of the least effective. Then, groups of judges and content analysts worked over the transcripts of the lectures, scoring and rating them on a host of variables. Some of these were sentence fragments, the average sentence length, the number of prepositional phrases per sentence, and so on. Other variables dealt with the number of self-references by the teacher, or with various aspects of syntax, or instructional set, familiarization, uses of previous knowledge, mobilizing sets, attention focusing procedures, organization, emphasis, amount of repetition and redundancy, the number of words per minute, and so on.

The variables that discriminated between the five best and the five worst lectures on Yugoslavia were then tried out on the other set of five best and five worst lectures on Yugoslavia to see if they still discriminated. Those that survived this first cross-validation were then tried out on the best and worst lectures on Thailand. At the last accounting, two characteristics of the lectures had survived this kind of validation and cross-validation procedure. These variables were what we are calling "explaining," or the degree to which the teacher describes the how, why, or effect of something, and the "rule-eg-rule" pattern, or the degree to which the teacher states a generalization, gives examples of it, and then summarizes a series of illustrations at a higher level of generality than the illustrations themselves. These two variables not only seem to be valid in our data but also are reliably rated by independent judges. Nonetheless, these must not be considered to be firmly established findings; they are merely examples of the kinds of conclusions to which research of this kind can lead.

Currently, we are in the process of scoring all of the explanations on all of the variables that appear to hold any promise, and we will then undertake studies of the complete correlation matrices involving not only the indices of explanation effectiveness, but also all of the characteristics of the explanations, and the ratings of the lectures by the pupils who heard them. Such a complete correlational study will throw light on the consistency from one lecture to another of the indices of lecture effectiveness, and the stylistic characteristics of the lectures, and also their intercorrelations.

What I have been describing is of course a correlational study. Along with its advantages in permitting the exploration of a wide variety of possible correlates of explaining ability as they occur under fairly normal conditions, it also has the disadvantage of making causal interpretations hazardous. For this reason, studies of this kind ought to proceed fairly rapidly into experiments in which the different ways of explaining will be based, at least in part, on leads obtained from our correlational studies.

Such experimental research may lead toward quite novel methods of teaching that could never be developed on the basis of studies of teaching the way teaching is. Stolurow (1965) has contrasted the approach of "modeling the master teacher" with that of "mastering the teaching model." The first approach is that of studying the most effective teachers we can find in order to find out how they behave and what they are like so that we can attempt to produce more teachers like them. Many research workers see little promise in this approach. They recommend that we undertake instead to develop wholly new models of the teaching process designed for optimal effectiveness regardless of their similarity to the way teaching now goes on in the normal classroom.

COMPUTER-ASSISTED PROGRAMMED INSTRUCTION

The teaching model that many advocates of this approach have in mind is that of programmed instruction, particularly computer-assisted programmed instruction. As Suppes and Atkinson and others have described this revolutionary undertaking in research and development on instruction, it holds out great promise indeed. Before too long, the annoying problems in the hardware will have been solved. After a somewhat longer time, we may expect substantial and well-validated programmed curricular materials to have been developed in all the subject matters and grade levels. As one who has seen the highly developed installations at the Brentwood School in East Palo Alto, California, I must share the optimism of Suppes and Atkinson, and other developers of computer-assisted instruction.

Their very success or coming success, raises problems for the kind of instructional methods with which we have been concerned in this paper. On superficial examination at least, certain major problems of ordinary classroom teaching seem to be clearly surmounted by computer-assisted instruction. For example, the problem of the cognitive complexity in the teacher's task, of how the teacher can say just the right thing at the right time to develop a concept or formulate a theory, is apparently well handled, at least in principle, by computer-assisted instruction. Its programs can be worked out and tried out in meticulous detail, well in advance, at leisure, by the most skilled curriculum experts in the land, and then made available in all their subtlety and complexity to every teacher who uses the program. Another major problem in the ordinary classroom is that of individualizing instruction. No matter how we group our pupils between schools, within schools, or within classrooms, we still have the problem of adjusting the rate and direction of the teaching and learning process to the needs and abilities of the individual pupil; here again, at least in principle, computer-assisted instruction seems at first glance to have the better of the live teacher.

While pondering these problems, I got some help from a restatement of the idea of individualized instruction in a recent paper by Philip Jackson (1966). He states it as follows:

> Individualizing instruction, in the educator's sense, means injecting humor into a lesson when a student seems to need it, and quickly becoming serious when he is ready to settle down to work; it means thinking of examples that are uniquely relevant to the student's previous experience and offering them at just the right time; it means feeling concerned over whether or not a student is progressing, and communicating that concern in a way that will be helpful; it means offering appropriate praise, not just because positive reinforcers strengthen response tendencies, but because the student's performance is deserving of human admiration; it means, in short, responding as an individual to an individual.

Individualization in this sense is much more than allowing for differences in speed of moving through a program or providing different branches or routes through the material.

Jackson's analysis of this kind of limitation in computer-assisted instruction should be placed alongside of the indications by Suppes (1966) that tutoring and dialog, which are higher levels of instruction than drill-practice, are still well in the future, as capabilities of computer-based instruction. Hence, any fears about the rapid obsolescence of live teachers, even where narrowly defined cognitive objectives are concerned, are quite unwarranted. That is, there will still be a need for teachers to use the kinds of technical skills, including explaining, with which the analytic approach being developed at Stanford and elsewhere is concerned. We shall have to continue to grapple with the problems of cognitive complexity and individualization through the medium of the live, human, teacher even in the realm of the well-formulated cognitive objectives. And there will always be the indispensable role of teachers in assisting pupils in attaining various kinds of affective and social learnings in the classroom.

Accordingly, group discussions, role playing, teaching for divergent thinking, as well as the technical skills I have already mentioned, are all the subject of various research and development projects now under way in the Stanford Center for Research and Development in Teaching. We are also looking at the way in which the organizational context influences the teacher's choices among ways to teach. And, in one of our projects, entitled "The Teacher in 1980," we are looking at the way in which new curriculum developments, television and other technical aids, computer technology, and new organizational schemes in the schools will affect the teacher's role in the foreseeable future.

CONCLUSION: A NEED FOR ANALYSIS

In conclusion, let me refer again to what I see as one basic new theme in the research and development in teaching that is now under way at Stanford and elsewhere. If it were necessary to sum it up in one word, my word would be analysis, breaking down the complexities that have proven to be so unmanageable when dealt with as a whole. We are no longer crippled by the notion that because there is one word called "teaching," there is one, single, over-all criterion of effectiveness in teaching that will take essentially the same form wherever teaching occurs. Even if none of the analyses of teaching that we have now proves to be viable, they will not be replaced by the old global conceptually impossible, complex variables that I see as the reason for the fruitlessness of so much of research on teaching in the past.

Instead, they will be replaced by other analyses of teaching, perhaps even finer analyses, until we get the sets of lawful relationships between variables that will mark the emergence of a scientific basis for the practice of teaching. It may well be that a 15-minute explanation of a 5-page magazine article is still too large a unit of teaching behavior to yield valid, lawful knowledge. It may well be that the mean score on a 10-item test of comprehension, adjusted for student ability and content relevance of the lecture, is still too large and complex a dependent variable. But, compared with the massive, tangled, and unanalyzable units that have typically been studied in the past – in research on the lecture method, the discussion method, and class size, for example – such units seem precise and manageable indeed. And eventually, of course, we shall have to put teaching back together again into syntheses that are better than the teaching that goes on now. I think it would be safe to say that there is now some hope of our being able to develop a scientifically grounded set of answers to every teacher's central question, "What should I do in the classroom?"

REFERENCES

Berelson, B. and Steiner, G. A. Human behavior: An inventory of scientific findings. New York: Harcourt, Brace, and World, 1964.

Fortune, J. C., Gage, N. L., and Shutes, R. E. Generality of the ability to explain. Paper presented at the meetings of the American Educational Research Association, Chicago, February, 1966.

Gage, N. L. Paradigms for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 94-141.

Jackson, P. W. The teacher and the machine: Observations on the impact of educational technology. Unpublished manuscript, University of Chicago, 1966.

Kuhn, T. S. The structure of scientific revolutions. Chicago: University of Chicago Press, 1962.

Rosenshine, B. Behavioral correlates of effectiveness in explaining. Unpublished doctoral dissertation, Stanford University, 1968.

Stolurow, L. M. Model the master teacher or master the teaching model. In J. D. Krumboltz (Ed.), Learning and the educational process. Chicago: Rand McNally, 1965. Pp. 223-247.

Suppes, P. The uses of computers in education. Scientific American, 1966, 215 (3), 207-220.

Walker, Helen M. Preface. In W. H. Lancelot, A. S. Barr, T. L. Torgerson, C. E. Johnson, V. E. Lyon, A. C. Walvoord, and G. L. Betts, The measurement of teaching efficiency. New York: Macmillan, 1935. Pp. ix-xiv.

Interaction Analysis and Inservice Training

Ned A. Flanders
The University of Michigan
Ann Arbor, Michigan

WHAT IS INTERACTION ANALYSIS?

INTERACTION analysis is a system for observing and coding the verbal interchange between a teacher and his pupils. The assumption is made that teaching behavior and pupil responses are expressed primarily through the spoken word as a series of verbal events which occur one after another. These events are identified, coded so as to preserve sequence, and tabulated systematically in order to represent a sample of the spontaneous teacher influence.

The most important criterion which any coding system must meet before it can be considered satisfactory is that a trained person can decode the data in order to reconstruct those aspects of the original behavior which were encoded, even though he was not present at the observation. A part of this article will describe inferences which can be made from a blind analysis of coded data.

Interaction analysis has been used to study spontaneous teaching behavior and it has also been used in projects which attempt to help teachers modify their behavior. In the first instance there may be a long period of time between observations and the analysis of the data. The data can be punched on IBM cards as they are collected over a period of several months, but a computer program to tabulate and analyze the data may be used only after all the observations have been completed. On the other hand, when interaction analysis data are collected in order to provide a teacher with information about his own behavior as a part of preservice or inservice training, then it may be advantageous to code directly into a desired tabulation form so that interpretations can be made at the earliest moment after the teaching episode is completed. This article will be more concerned with the procedures of interaction analysis which can be used during preservice and inservice training, and less concerned with applications in more basic research projects.

A possible goal of an inservice training program might be to discover whether the spontaneous patterns of verbal communication which are observed are, or are not, consistent with the intentions of the teacher. In such a program, the assumption is made that modifying behavior in order to make it more consistent with intent will, in most instances, result in an improvement. The model toward which behavior is modified is created by the teacher attempting to change.

An inservice training program can also be designed in which all participating teachers attempt to make similar modifications, for example, developing more skill in making full use of the ideas which are expressed by pupils during classroom discourse. Here the model could be justified from the results of research which made use of interaction analysis, but special care would be necessary in designing activities which would help teachers accept and understand the desirability of such modifications. In brief, the value orientation would be that the quality of classroom instruction is improved when the ideas expressed by pupils are more adequately recognized, clarified, used in some step of a logical analysis, thus giving the pupils a more active part in the learning activities.

In nearly all applications of interaction analysis it is desirable to collect other kinds of data such as pupil attitudes, pupil achievement, and perceptions of the teaching situation held by the teacher and the pupils. These additional data permit the development of theory and explanation.

The resources of such a program, as with all inservice activities, require money in order to obtain time, space, and assistance for those who are participating. A number of steps can be anticipated: (a) Prospective observers and participating teachers must be trained to use the technique. This usually requires 6 to 12 hours under the direction of a qualified observer. (b) Personal, self-development goals must be clarified by the participating teachers. (c) Each teacher-observer team must set up comparison situations involving a planned change of behavior and two or more observations should be scheduled so that evidence of change can be assessed. (d) Plans to collect other types of information must be completed. (e) Social-skill training sessions, based on interaction analysis categories, are often helpful, especially when they are incorporated in the observation training. Such training often helps teachers set personal goals for changing their behavior. (f) Plans will be necessary to provide incentives for teachers and to maintain the momentum of the program, once it is started. (g) Care must be taken to insure freedom from threat, to make sure that the voluntary aspects of participation are genuine, and to avoid superimposing the program as an excessive demand on teachers who may already be too busy.

OBSERVATION PROCEDURES AND MATRIX INTERPRETATION

Given ten categories, shown in Table 1, all verbal statements are classified at least once every 3 seconds by a trained observer. The events are coded by using the arabic numbers from one to ten which are written down in such a way as to preserve the original sequence. The data can then bé tabulated in a table of ten rows and ten columns which is called a matrix.

Such a series is entered into a matrix two at a time. The first number of each pair indicates the row of the matrix, the second the column. The first pair consists of the first two numbers. The second pair consists of the second and third numbers, and thus overlaps the first pair. All tallies enter the matrix as a series of overlapping pairs.

With one tally approximately every 3 seconds, there are one hundred tallies for 5 minutes, 1, 200 tallies per hour; therefore, 20 minutes, or about four hundred tallies, provide a matrix with sufficient data for a number of inferences about verbal communication.

In a sustained observation of a teacher covering six to eight 1-hour visits, it is necessary to tabulate separate matrices for different types of classroom activities. Each matrix should represent either a single episode of class activity or any number of homogeneous episodes that are combined. We use five activity categories for junior high school academic subjects; they are: routine procedures, discussion of new material, discussion to evaluate student performance or products of learning, general discussion, and the supervision of seatwork or group activities. Different activity categories may be useful for a self-contained elementary-school classroom. In any case, the purpose of tabulating the data in several different matrices, instead of in just one total matrix, is determined by the purposes of observation and the range of expected classroom activities.

In the case of the inservice training of teachers, data are tabulated into separate matrices in such a way as not to mask the comparison to be made. To illustrate, suppose the comparison is between two samples of teaching behavior, one before and one after several weeks of social skill training. Keeping the data in homogeneous activity matrices will help to avoid false conclusions. For example, a decrease, increase, or no change in certain categories, when the two matrices are compared, may be due to the differences in the learning activities observed, rather than due to change in teaching behavior which resulted from inservice training. Grouping the data to represent homogeneous learning activities helps to avoid such confusion.

A tabulated matrix divides into special areas for interpretations that are shown in Table 2. Particular questions can be answered by comparing tallies within and between these areas. Here are some examples.

Areas A (1+2+3+4), B(5+6+7), C(8+9), and D(10) can be used to find the percent time the teacher talks, the pupils talk, and time spent in pauses, silence, and confusion. Comparisons between Areas A and B provide information about the relative balance between initiating and responding within teacher talk. Initiating teacher talk is more directive, tends to support the use of teacher authority, and restricts pupil participation. Responsive teacher talk is more indirect, tends to share authority, and expands pupil participation.

Area E is a block of nine cells that indicates the continued use of acceptance and praise, constructive reaction to pupil feeling, and clarifying, accepting, and developing pupil ideas, as well as transitions among these three categories while the teacher is talking. In any inservice training program devoted to increasing the teacher's attention to ideas expressed by pupils, before and after comparisons would require an analysis of these nine cells. In fact, an inservice training program which attempted to teach more subtle differences in the teacher's reaction to pupil ideas might require subdividing Category 3 in order to note the presence and absence of various types of Category 3 statements. For example, 3-1 -- merely repeats to show that the pupil ideas were heard; 3-2 -- reacts to specific pupil ideas, but only in terms of the teacher's perceptions of these ideas; 3-3 -- reacts to specific pupil ideas, but reactions incorporate the perceptions of one or more pupils; and 3-4 -- stimulate a reaction to a pupil's ideas by asking questions so

TABLE 1

CATEGORIES OF INTERACTION ANALYSIS

Teacher Talk and Student Talk

TEACHER TALK

Response

1. * Accepts feeling: accepts and clarifies the feeling tone of the students in a non-threatening manner. Feelings may be positive or negative. Predicting or recalling feelings are included.
2. * Praises or encourages: praises or encourages student action or behavior. Jokes that release tension, but at the expense of another individual; nodding head, or saying "um hm?" or "go on" are included.
3. * Accepts or uses ideas of students: clarifying, building, or developing ideas suggested by a student. As teacher brings more of his own ideas into play, shift to category five.
4. * Asks questions: asking a question about content or procedure with the intent that a student answer.

Initiation

5. * Lecturing: giving facts or opinions about content or procedures; expressing his own ideas, asking rhetorical questions.
6. * Giving directions: directions, commands, or orders to which a student is expected to comply.
7. * Criticizing or justifying authority: statements intended to change student behavior from non-acceptable to acceptable pattern; bawling someone out; stating why the teacher is doing what he is doing; extreme self-reference.

STUDENT TALK

Response

8. * Student talk – response: talk by students in response to teacher. Teacher initiates the contact or solicits student statement.

Initiation

9. * Student talk – initiation: talk by students which they initiate. If "calling on" student is only to indicate who may talk next, observer must decide whether student wanted to talk. If he did, use this category.
10. * Silence or confusion: pauses, short periods of silence and periods of confusion in which communication cannot be understood by the observer.

*There is NO scale implied by these numbers. Each number is classificatory; it designates a particular kind of communication event. To write these numbers down during observation is to enumerate, not to judge a position on a scale.

that other pupils react. In effect, Category 3 is expanded into four categories for a special purpose. This would result in a 13 x 13 matrix instead of a 10 x 10 matrix.

Area F is a block of four cells that indicates the continued use of directions and criticism and transitions between these two categories. The two transition cells are particularly reliable indicators of discipline problems. Shifting from directions to criticism is tallied in the 6-7 cell, and indicates that expected compliance is judged unsatisfactory by the teacher. Shifting from criticism back to directions, the 7-6 cell, indicates a return to more directions after criticism.

Areas G_1 and G_2 are particularly interesting because they isolate the immediate response of the teacher at the moment students stop talking. One aspect of teacher flexibility can be discovered by comparing the balance of indirect and direct statements shown in G_1 and G_2 with those found in Areas A and B. The difference between superficial, short, perfunctory praise or clarification, and praise or clarification that is more carefully developed is easily seen by comparing the tallies in Area G_1 with those in E, particularly the 2-2 and 3-3 cells.

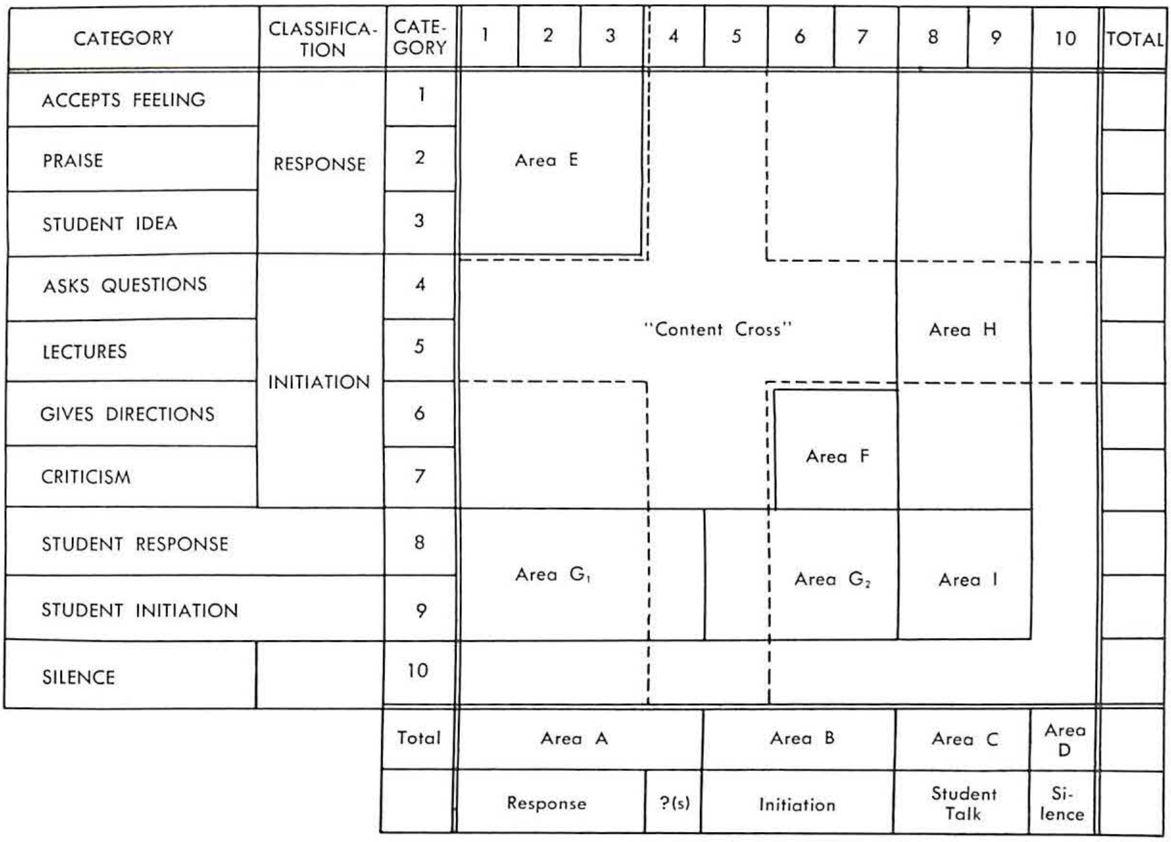
Area H indicates the types of teacher statements that trigger student participation. Responses to the teacher are found in column 8; statements initiated by the student in column 9. As one might expect, there is usually a heavy loading of tallies in the 4-8 cell. High frequencies in this cell and the 8-4 cell, but not in the 8-8 cell, often indicate rapid drill.

Area I indicates sustained student participation. These may be lengthy statements by a few students, or student-to student communication.

So-called "steady state" cells fall on the diagonal from cell 1-1 to 10-10. Tallies here indicate

TABLE 2

AREAS OF MATRIX ANALYSIS



CATEGORY	CLASSIFICATION	CATEGORY	1	2	3	4	5	6	7	8	9	10	TOTAL
ACCEPTS FEELING	RESPONSE	1											
PRAISE		2		Area E									
STUDENT IDEA		3											
ASKS QUESTIONS	INITIATION	4											
LECTURES		5					"Content Cross"				Area H		
GIVES DIRECTIONS		6											
CRITICISM		7							Area F				
STUDENT RESPONSE		8											
STUDENT INITIATION		9		Area G₁				Area G₂			Area I		
SILENCE		10											
		Total	Area A				Area B			Area C		Area D	
			Response			?(s)	Initiation			Student Talk		Silence	

that the speaker persists in a particular communication category for longer than 3 seconds. All other cells are transition cells moving from one category to another.

Outlined in the center of Table 2 by dash lines is the content cross. The total number of tallies in this area, compared with tallies not in this area, gives a very crude indication of the content orientation of the class activity.

In addition to making use of the areas just described, the following procedure can be followed to interpret a matrix.

First, locate the single cell within the ten rows and columns which has the highest frequency. The pair of events, represented by the cell, is the most frequently occurring and can be used as a starting point in reconstructing the interaction.

Second, from this highest frequency cell, you start forward or backward, in terms of sequence, to begin a sequence diagram. The row of any cell indicates the most likely third event, that is, the event which is most likely to follow, given an original pair of events designed by the highest frequency cell. The column, on the other hand, indicates which event most probably preceeded the pair of events in question. The flow of events is properly represented when the eye scans the matrix in a clockwise rotation. Should the highest frequency fall into a transition cell, not a steady state cell, the row or column of either number in the pair can be studied to retrace or advance the sequence of events.

An example of matrix interpretation will be shown later in this article. Skill in matrix interpretation, however, is not likely to develop from reading this article, which serves only to propose guidelines. For that matter, skill in observation cannot develop from reading about how it is done. All aspects of interaction analysis require practice in order to develop skill. It is the opposite of a spectator sport.

FEEDBACK AND CONSULTATION WITH TEACHERS

The purpose of feedback is to provide a teacher with information about his verbal statements which permits a comparison with some standard or model of what should have happened. Without purposive comparisons which are planned in advance of the observation, the reactions of the person receiving the information may be reduced to incidental speculations or points of interest that happen to be noticed.

To be useful as a model or standard, one's intentions must be specified in terms of frequencies to be found in the cells of the matrix. Thus, if a teacher wishes to practice providing more extended praise, he will expect to find an increase in the 2-2 cell of the matrix, one situation compared to another. If a teacher decides he would like to stimulate more pupil talk in which ideas are initiated by pupils, he might study column 9 in the matrix to see what events trigger these pupil statements. In a program of inservice training in which all teachers attempt to increase the utilization of ideas expressed by pupils, attention will be directed to column 3 and row 3 of the matrix.

Most inservice training programs can achieve some success in bringing selected concepts and value orientations to the attention of teachers, these are matters of awareness. Evidence that such awareness has been implemented through overt behavior requires an objective assessment of spontaneous teaching behavior.

The assessment of the spontaneous behavior must be reasonably objective in order to be reliable. Unfortunately, interaction analysis is not free of bias and error, probably about one out of every ten classifications of an experienced observer is incorrect. Interaction analysis data can be and probably are more objective, when dealt with in summary form, compared with most other procedures for making judgments about spontaneous teaching behavior. Judgments about events which occur within time segments of only a few seconds and which must be repeated again and again tend to become more consistent with practice. Furthermore, noting the presence or absence of a short event is not a procedure which lends itself as easily to distortion and bias.

No matter how objective, reliable, and valid an assessment procedure, the results will be distorted if the behavior itself is distorted. Unfortunately, merely anticipating observation might cause non-representative behavior to appear, not to mention the observation experience itself. Below are some policies and suggestions which we have found helpful in reducing the tendency of a teacher to put on an act while being observed.

First, an observer should be in the classroom only when invited by the teacher.

Second, the invitation should be based on a plan of inquiry which was developed by the teacher and observer prior to any classroom visits. Observation should produce information which is relevant to some problem or question which is considered important to both participants. Thus, a teacher participating in an inservice training program which proposes to improve the way pupil ideas are handled during classroom discussion may be curious about this aspect of interaction before and after training. Such a question might involve creating two similar lesson plans in which a teacher would be confronted with opportunities to react to pupil ideas. One lesson would be observed before training and the second lesson after training. The plan could be embellished to provide greater insights by collecting additional data. For example, predictions about pupil perceptions, teacher perceptions, pupil attitudes and similar phenomena could be made, one lesson compared with the other. Then, instead of merely counting the incidence of constructive teacher reactions to ideas expressed by pupils, certain theories about the consequences of such teacher behavior might be investigated.

Third, the status and power difference between the observer and the teacher should be a minimum. Another teacher who is a best friend might make the most appropriate observer, providing skill in observation is present.

Fourth, the conference to provide feedback should follow a logical plan of inquiry. All the relevant data should be at hand and referred to in terms of questions to be answered and not in terms of idle curiosity. It often helps to have two or more matrices, since this facilitates the making of comparisons out of which theoretical explanation grows. A single matrix more often stimulates opinions about what is "good" and "bad," illustrated by the question, "Do you think the lesson was satisfactory?" Such questions place the observer in an awkward position, since he must express a general judgement. In this position he cannot be an equal partner in the inquiry process. On the other hand, when a hypothesis about behavior is being investigated and professional competence of the teacher is not in the foreground, the analysis of data is more systematic and unwanted defensive distractions are less likely to occur.

Fifth, the entire procedure including planning, execution, and analysis usually works more smoothly when the teacher, as well as the observer, has had approximately equal experience in observation. For example, setting up hypotheses, designing two comparable lesson plans, and knowing where to look in a matrix for the proper information are phases of the experience which should be shared by two partners who are equally competent. When the observer is more experienced and competent, the teacher defers and becomes the dependent member of the team. Both members of the team should have had previous experience in both teaching and observation.

ILLUSTRATIONS OF PRACTICAL PROCEDURES

For purposes of illustration, let us assume that the goal of inservice training is to increase the teacher's skill in making use of ideas expressed by pupils. In an article as short as this one, only four aspects of such an inservice training program will be mentioned. First, some initial performance data provides a before training performance pattern for which subsequent observation data can be compared. Second, skill training procedures can be closely correlated to observation procedures. Third, data in

TABLE 3

OBSERVATION MATRIX

CATEGORY	1	2	3	4	5	6	7	8	9	10	Total
1	–	–	–	–	–	–	–	–	–	–	–
2	–	1	1	1	2	–	–	1	5	–	11
3	–	–	5	1	4	–	–	–	–	–	10
4	–	–	–	23	2	1	–	42	3	5	76
5	–	2	1	22	80	1	2	3	3	3	117
6	–	–	–	1	–	–	1	3	–	–	5
7	–	–	–	–	2	1	1	–	–	–	4
8	–	5	–	22	19	–	–	45	7	–	98
9	–	3	3	3	7	–	–	3	32	–	51
10	–	–	–	3	1	2	–	1	1	–	8
Total	–	11	10	76	117	5	4	98	51	8	380
%	–	2.9	2.6	20.0	30.8	1.3	1.1	25.8	13.4	2.1	100.
of	25.5				33.2			39.2		2.1	
Total	Teacher Total: 58.7							Student Total		Silence	

I/D = 0.77 Content Cross = 70.8%

$I/D_{8,9}$ = 1.38 Steady State = 49 %

addition to interaction analysis are helpful in deciding whether or not a change in behavior is an improvement. And fourth, more advanced training designs and more complex data collection procedures will be necessary to push progress beyond the initial results. Each of these four topics will be discussed in turn.

INITIAL PERFORMANCE DATA

Initial performance data might be in the form of a short observation during a class discussion. Table 3 shows a matrix of a teacher which will now be interpreted to show how the observer and teacher make an initial diagnosis before training. The same data, of course, can serve as a before training standard in order to determine whether change has occurred.

Since the total tallies equals 380, one can estimate that the matrix represents about 19 minutes of interaction (100 tallies = 5 minutes, 20 tallies = 1 minute).

A number of percentages and other ratios, which help to form an initial picture, can be found at the bottom rows of Table 3. For example, the teacher talked 58.7 percent, the pupils 39.2 percent, and silence and confusion was 2.1 percent. The teacher was fairly directive, that is he initiated more than he responded, as shown by an I/D = 0.77 (divide all tallies in categories 1 + 2 + 3 + 4 by 5 + 6 + 7 to obtain this ratio). The highest cell frequencies are found in the steady state diagonal cells, such as the 5-5, the 8-8, and the 9-9 cells. This suggests that the teacher and pupils were able to continue a particular mode of expression once it started. Higher frequencies in these steady state cells indicate that the tempo of exchange was slower, for example, than might occur in a drill period. All silences were 3 seconds or less (note no tallies in the 10-10 cell) and most pauses followed teacher questions (N = 5 in the 4-10 cell) and teacher lecture (N = 3 in the 5-10 cell) rather than pupil statements. In six of these eight transitions, it was the teacher who broke the pause by talking (note that N = 1 in the 10-8 and 10-9 cells). This analysis of silence supports the interpretation that the teacher tended to initiate and did not permit a pupil more than 3 seconds to respond.

The matrix is sometimes more easily understood when it is translated into a flow pattern illustrated in Figure 1. In this diagram the most frequently occurring steady state cells are shown as rectangles and the size of the rectangle indicates the relative frequency of the pair. Transitions among these cells are indicated by arrows and the thickness of the arrow is roughly proportional to the frequency of these transitions. Anyone can learn to draw such a pattern flow diagram from a tabulated matrix. Begin with the highest frequency cell, in this case the

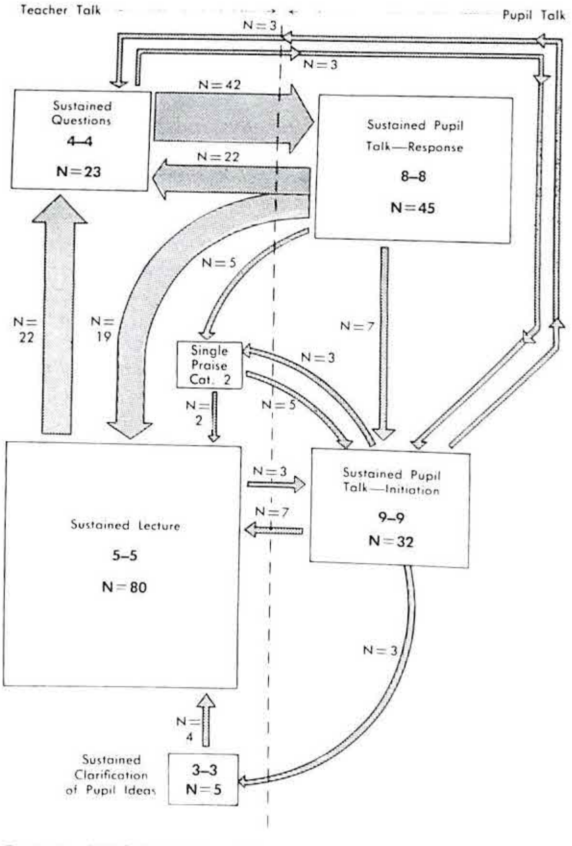


Figure 1: Initial Flow Pattern Before Training

5-5 cell with eighty tallies. Proceed across the same row to find out the next most likely occurring events and inspect columns to identify the most likely preceding events. The arrows in Table 3 illustrate the clockwise direction of the flow. Thus in both the matrix and in Figure 1, begin with the 5-5 cell, sustained lecturing. The next most likely event is that the teacher will ask a question. Next, pupils are most likely to respond to this question, see category 8. There are two events of almost equal probability following pupil response talk, one is that the teacher will ask another question and the other is that the teacher will lecture. The four transitions shown by the heavy arrows in Figure 1 account for most of the interaction: these are, 5-4, 4-8, 8-4, and 8-5. Other transitions do occur, but the transitions just listed and their associated steady state cells account for 285 tallies or about 75 percent of the entire observation.

Many things could be said about this pattern of interaction, but by far the most certain inference is that in the event that this teacher increased his use of category 3, it could be clearly seen in the matrix. The use of category 3, especially the 3-3 cell, is below average for a classroom teacher. After training along these lines a second matrix and flow pattern diagram could be made as a comparison with this initial performance.

SPECIFIC SOCIAL SKILL TRAINING EXERCISES

Specific social skill training exercises can be conducted which are closely correlated with interaction analysis. A training skill exercise might involve three or four people who take turns providing certain patterns of verbal statements while responding spontaneously to the statements of others in the group. The person being trained is designated as the actor, one other person is the observer, and the rest are reactors or foils who provide spontaneity. Different assignments are given to the actor which he then carries out by interacting with his foils. The observer keeps a record of the communication by coding within the ten categories. The assignment can be given in terms of the categories, for example, the actor is asked to produce a 4-8-4-8-4-8-4-8, etc., sequence. An actor might be asked to choose a topic and produce any sequence which is relevant to teaching and is related to the goals of the training.

The activity of a training exercise follows a model in which an actor attempts a particular pattern, his foils provide a spontaneous setting, and the observer keeps a coded record of what happens. Some training episodes may last only a few minutes, others are longer. Each episode is followed by a discussion of the performance and ways that it can be improved. While this kind of activity can easily degenerate into superficial play acting, especially when first attempted, it is also true that serious and interested participants can use this technique to practice producing certain patterns of behavior under increasing difficult circumstances and make progress in the self control of spontaneous behavior.

Learning how to make full use of the ideas expressed by pupils is a goal to which spontaneous skill training exercises can easily be adapted. It is possible to gain practice in listening carefully, in listing ideas expressed by another person, in summarizing ideas which have just been expressed, in building questions on ideas expressed by others, and to teach pupils to initiate their own steps in problem solving by showing a pupil how his ideas are related to a problem.

Skill practice sessions become more realistic when they are closely related to classroom practice sessions. After one or two initial observations an observer and teacher may discuss the relationships between the teacher's intention and the patterns which appear on several matrices. Skill practice sessions are then designed to emphasize a personal goal of the teacher based on discrepancies revealed by the matrix. If the training activities are custom built to the needs of the teacher, and if he sees them as relevant to his own professional development, then he is more likely to approach the training sessions with serious intent and a sense of optimism concerning personal development. It also helps when classroom patterns begin to change in a desired direction as a result of skill practice sessions.

Usually, the assignments of actor, foil, and observer are rotated during skill practice sessions. In this way a teacher not only becomes familiar with the coding system of interaction analysis, he also obtains some degree of observer proficiency. He also learns to accept and develop the perceptions that others have of his behavior.

The primary purpose of these training episodes is to practice producing certain patterns of statements, to translate concepts about teaching into spontaneous behavior patterns, to learn how to assess such patterns, and to discuss how these various patterns appear in classroom teaching. Many variations of training exercises can be directed toward these goals making them useful for inservice training.

DECIDING WHEN A CHANGE IS AN IMPROVEMENT

Deciding when a change is an improvement usually requires data in addition to interaction analysis. For example, merely giving a teacher instructions to "go in there and produce more threes," much like a football coach, is quite likely to increase the incidence of category three. This change may take place, however, without any insight into the teaching process and may possibly be seen as either inappropriate or not realistic by the pupils.

Utilizing the ideas expressed by pupils during classroom discourse involves several pedagogically sound principles. One such principle is that this kind of teacher behavior stimulates the perception among pupils that each pupil is free to express his ideas. Simple paper and pencil reaction sheets, to be filled out by pupils old enough to read and write, can be used to assess pupil perceptions immediately following a practice session in which the teacher tries to accept and clarify pupil ideas. When both the observer's records and the pupil's average perceptions from a paper and pencil instrument indicate greater expression and use of pupil ideas, then the additional evidence provides greater confidence that the change is an improvement.

Another pedagogical principle is that a pupil learns to cluster ideas because they are similar and then abstract the cluster with an appropriate label through directed practice in expressing his ideas. A second observer can keep an inventory of separate ideas expressed by pupils, he can note instances in which the pupils, instead of the teacher, noticed a cluster, and finally, he can record whether the teacher or the pupils supplied an appropriate label to a given cluster. This additional evidence also helps to show whether a change in teacher behavior is, or is not, an improvement.

PUSHING TRAINING BEYOND SUPERFICIAL CHANGE

Pushing training beyond superficial change usually requires special category systems of interaction analysis. For example, the division of category 3, mentioned earlier, can provide a record of different kinds of teacher reactions to the ideas expressed by pupils. A similar expansion of categories 8 and 9 will show variations in the different kinds of pupil statements. Often the expansion of teacher talk categories can provide more intricate and difficult social skill training assignments. Working back and forth between more elaborate category systems and more complex spontaneous social skill training assignments helps a teacher understand principles of pedagogy in terms of his own behavior. Unless a teacher can act out his insights about teaching, these insights are of little use for the improvement of instruction.

SUMMARY

The use of interaction analysis in an inservice training program places an emphasis on the analysis of spontaneous verbal behavior. This emphasis helps to translate ideas about teaching into classroom application.

For those who would like to try classifying verbal statements, but are inexperienced, a number of references are listed at the end of this article which include suggestions for beginning.

REFERENCES

Flanders, N. A. Intent, action and feedback: A preparation for teaching. Journal of Teacher Education, 1963, 14, 251-260.

Flanders, N. A. The role of the teacher in the classroom. Minneapolis, Minnesota: Amidon, 1963.

Flanders, N. A. Interaction analysis in the classroom: A manual for observers. (Rev. ed.) School of Education, University of Michigan, 1965.

Flanders, N. A. Teacher influence, pupil attitudes, and achievement. Cooperative Research Monograph No. 12, U. S. Office Of Education, OE-25040. Washington, D. C.: U. S. Government Printing Office, 1965.

PART V

The Current Scene

THE CURRENT scene in educational research and development is one of rapid and unprecedented change brought about largely by the advent of Federal funding. In the first paper in Part V, R. Louis Bright and Hendrik D. Gideonse trace the development of these changes with reference to involvement of the Federal government starting with the 1954 legislation. They cite the national need for the results of educational research and development as the basis for the decision of the government to provide grant dollars for developmental projects and the design of the necessary strategies to implement programs of research and development. The evolving strategies of research and development designed to improve education are delineated by the authors and include 1) curriculum development, 2) research and development centers, 3) the educational research information centers, 4) a program for training educational researchers, and 5) a system of autonomous educational laboratories. These strategies have developed rapidly and promise to provide the basis for much of the future growth of educational research and development.

The costs of educational research and development have grown from the original 1.7 million appropriated in 1957 to the present level of approximately 200 million. This represents less than one-half of one percent of the total amount expended for education by all phases of our society, a figure which is far under the standards for research and development investments in most major industries. Clearly additional growth can be anticipated. The authors of this paper propose additional major expenditures in research and development amounting to some billion dollars annually which would represent the still modest figure of approximately 2.2 percent of the total expenditures for education.

This continued growth of research and development in education, which is almost exponential, calls for careful and systematic planning. In answer to this need the Bureau of Research, U. S. Office of Education, has proposed the establishment of Educational Policy Centers to conduct studies for establishing the means and directions of educational research and development 5, 10, 20, and 30 years hence. The rationale of these policy centers is discussed by Bright and Gideonse.

The rationale and focus of a specific Research and Development Center is presented by Herbert J. Klausmeier, The Director of the Wisconsin Research and Development Center for Cognitive Learning. The Research and Development Center concept was proposed by the U. S. Office of Education in an effort to better coordinate research and development activities and to establish programatic research and development in specific areas related to the educational process. The Wisconsin Center is described in terms of eight components of operation: (1) problem area, (2) outcomes sought, (3) research and development program and strategy, (4) dissemination and adoption program and strategy, (5) staff, (6) administration and management, (7) locus of operation, and (8) source of stability of funding. Klausmeier points out that the programatic activities of the Wisconsin Center which have been brought sharply into focus now include 1) processes and conditions of learning, 2) processes and programs of instruction, 3) facilitated environments. Other programs deal with dissemination, training activities, and technical support. Such a center concept brings to bear on problems of education the diverse talents of the interdisciplinary staff available at a major American university.

Most new and complex activities are more readily understood if represented in a model idealizing their essential components and operations. The total research and development effort in education is no exception. The model proposed by Hendrik Gideonse in the final paper recognizes the independent characteristics of the research act and the development act. In a prior model proposed by Guba and Clark the research activity precedes and feeds results into the development activity which in turn provides materials and programs for adoption. However, Gideonse proposes a model which emphasizes the nonlinearity of the research and development complex. He states, "My purpose in developing an alternative model is to create a heuristic which illustrates the essential differences between research and development activities and show how the two are—or can be—related to one another and to the operating educational system."

The model proposed emphasizes the quite different objectives and outputs of the research and development activities. Research is characterized as an activity begun with specific outcomes not being known. Contrasted to this is the development activity which is begun with the objective known or established at the onset. In the model proposed the dynamics of initiation of research and development activities are quite different, as are the outcomes or products. Research may be initiated for its own sake and pursued solely for the knowledge which it produces. Whereas development activities are initiated in response to a specific need; the specifications for the product can frequently be defined before the onset of the development activity. However, even though the utilization of the research product is not predicted or predictable there rests on the researcher the obligation to present results in such a way so as to allow transfer to other research or development projects. In the model proposed by Gideonse the output function of the various activities is crucial. Such a model allows those concerned with planning research and development activities to more clearly see the essential characteristics of these activities.

The authors of this section describe the current scene and provide a glimpse of the future of educational research and development. The growth of the present research and development investment may indeed be a prologue to the future.

Research and Development Strategies: The Current Scene

R. Louis Bright and Hendrik D. Gideonse
United States Office of Education
Washington, D. C.

IT IS MOST stimulating to participate in a field which has generated as much excitement as educational research has in recent months. Controversy, discussion, and debate are very much the order of the day; the gradual growth of contacts with new and informative sources of data, and of faith in the research effort are developments full of promise for the future.

In agreeing to prepare this paper, we explicitly accept the notion that it makes sense to consider strategies of research and development for improving education. At the same time we do not make any claim to producing anything more than tentative steps in the direction of more adequately defined strategies for achieving mutually agreed upon objectives. For a variety of reasons which we will attempt to spell out in this paper, we have begun the process of developing and applying more explicit strategies to the research and development process. These strategies are neither rigid nor lightly held. Some notions of what we are about and why are clearly necessary; not to have them is to default a major responsibility. In this paper we try to spell out both the general outlines of the strategies we are pursuing and the procedures by which we hope to continually modify the specific substance implicit in those strategies and the actual strategies themselves.

The improvement of American education depends upon the systematic investigation of the process and the necessary conditions for learning, the development of instructional objectives, strategies, and materials based on the knowledge educators and others accumulate about the learning process, and finally on the implementation of those strategies and the use of those materials in instructional settings across the country. We take this as a fundamental article of faith, but it is a faith which has been substantiated over and over again in all fields of human endeavor whether they relate to space, health, poverty, industry or education. To claim, however, that significant and continuing advances in the process of education depend upon research and development is only the barest of opening statements. A perspective and the setting need first to be sketched in, and then a full exploration of the ramifications of that article of faith needs to be conducted.

SETTING THE PERSPECTIVE

Any consideration of the role of research and development in American education necessarily relates to the nature of the existing "system" through which we have institutionalized education in our society. With legal responsibility for education in this country vested in the states and operational responsibility for schools largely delegated to some 25,000 local educational agencies authorized by the states, the role of the Federal government has largely been confined to providing specified categorical financial support.

These different levels of responsibility are complicated by another feature of American education. Not many of the diverse institutions and agencies which serve education in this society are well-coordinated or formally related to one another. When we use the word "system" as applied to American education, we are being generous. What we have is a de facto system which serves the broad needs of the society well, but perhaps not as well as it might if it were generously laced with efficient communication channels designed to make it an integrated, functionally related entity of many parts and many purposes.

From the point of view of the general improvement of education, it makes more sense to talk of the educational system as composed of classes of institutions and agencies. Each class embraces organizations which perform similar functions and which, if there were an organized system, would be expected to fulfill a particular role. Thus, the classes of institutions include the three levels of responsibility for education – local, state, and federal. They include professional associations and institutions for training teachers. They include the industries which supply materials and equipment to the schools. They include universities and institutes under whose aegis research on education and learning is performed. The existing "system," in other words, is largely a collection of nearly unrelated sub-systems or institutions representative of sub-systems. This feature is of no small importance in devising strategies for the general improvement of the whole.

A NATIONWIDE NEED FOR EDUCATIONAL RESEARCH AND DEVELOPMENT

The nationwide nature of the need for educational research and development is becoming increasingly apparent. Schools, districts, state education agencies, private institutions, colleges, and universities all have major needs relating to the improved fulfillment of their instructional responsibilities. While individual institutions and agencies differ, many if not a majority of the problems which confront them are similar. Serious questions of economics of scale have been answered in effect by the relatively small amounts of money which have been devoted to educational research at the local level. We have seen for decades an entirely reasonable response on the part of local and state agencies with respect to the funding of major research and development efforts. While they have no doubt been aware that sophisticated, well-conceived research programs would benefit not only themselves but also many others as well, very few have felt that they themselves could assume the financial and human burden of mounting large-scale research programs. It would make vastly more sense if the actual cost of generating those findings and materials were amortized more proportionately across all the potential beneficiaries.

The method for teaching mathematics, for example, is not likely to be much different for similar children in Maine and Arizona. On the other hand, there may be significant differences in instruction in mathematics within a state for children having different socio-economic characteristics, learning styles, or motivational problems. It would be a philanthropic district or state that felt it could mount the requisite research and development projects to serve all the immediate local needs when some other mechanisms could be developed which would distribute the cost more fairly in return for the mutual benefits that would be derived.

THE EMPLOYMENT OF NATIONAL RESOURCES

The resources for conducting educational research and development, though not as limited as some people would have us believe, are in fact unequally distributed across the country. No necessary conclusions with respect to continuing patterns of supporting educational research can be drawn from that fact, but a consideration of it is certainly one of the relevant contextual matters for the development of research and development strategies. The nationwide character of the need for research and the unequal distribution of the human resources for research reinforce one another. They highlight the breadth of the problem. In other words, so long as human resources are unequally distributed, the direct or implicit coordination of them to serve research and educational needs of high priority will constitute one of the major background considerations in research strategies.

FINANCIAL RESOURCES

Certainly one of the more important matters lending perspective to research and development (R and D) strategies for improving education is the availability of financing. The resources are limited. Money available from all sources for R and D in education has never exceeded 1 percent of the total annual expenditures on education in the nation as a whole. Estimated expenditures for this fiscal year for all levels of formal education are $48.8 billion.[1] By contrast the funds available at the Federal level for educational research and development are somewhere between $200-220 million.[2] Simple arithmetic reveals that this amount constitutes less than half of 1 percent of the total annual expenditure on education. We do not by any means take that as "a given" for the future, but it is most assuredly an important element in developing perspective on the problem of research strategies for the present.

THE DISTINCTION BETWEEN "RESEARCH" AND "IMPROVEMENT"

A final element for consideration in this context is our conviction that R and D in education and the change process in education are two very different functions or processes which must, however, be related intimately, sensitively, and integrally to one another. The research and development process produces the substance of improvement. It yields the knowledge about the learning process and the social setting for learning and education, and encourages the refinement of objectives and the development of materials and instructional processes which incorporate that knowledge into improved educational practices.

The implementation of those practices in the school systems across the nation – the differential accomodation of different institutions and agencies to the new roles and expectations created by the improvements – however, is a very different kind of problem and requires strategies and tactics commensurate to the task as it is revealed. The differences between the two processes and the unique characteristics of each may very well shape significantly the strategies adopted for one or the other. The functional demands of educational research and development, for example, place certain constraints on the process of implementation. Similarly, the exigencies of the process of implementation may impinge upon the strategies for R and D which may be adopted.

RESEARCH, DEVELOPMENT, AND THE PROCESS OF IMPLEMENTATION

Recent policy-level discussions of educational research and development within agencies responsible for research funds, between such agencies and their constituencies, and within the research community itself, have made it clear that there is far from unanimous agreement on the definition of the terms commonly applied to research and development. For this reason, we believe it would be helpful to establish how we are now using various terms – not to develop for all time a thesaurus as applied to the research process, but rather to clarify our concepts and interpretations.

The research and development process involves all those activities designed to produce knowledge about the learning process (research, for example, designed to improve learning theory or to improve our understanding of the fundamental chemical, biological, and neurological processes underlying learning). It includes other kinds of activities devoted to the social context of learning (for example, the classroom or school as a social system, or the relationship among non-instructional factors such as motivation of peers, socio-economic status, educational level of the home and the direct instructional interventions of the school such as curriculum, instructional methodology, facilities and so on). It includes studies pertaining to the organization and administration of the instructional enterprise.

Research and development also includes the development of new practices, materials, and processes based on the knowledge newly created in the other kinds of studies. This involves identification of the specific improvement desired and then the systematic engineering of the improvement, using all the information derived from basic research that is relevant.

A third part of the process, closely related to development, is the demonstration of the feasibility of the new materials, processes, or organizational forms in other than the hothouse setting of the development laboratory.

Finally, a fourth element in the R and D process is the systematic dissemination of information relating to all the different stages. Information about research projects completed and underway is essential to the basic researcher. The educational developer also needs to know the latest findings about learning if he is to be able to incorporate them in his work. Researchers and educators in the schools need to know what the development people are up to. On the other hand, the kinds of problems encountered by those engaged in curriculum or hardware development may suggest research projects to the psychologists and sociologists. School people need to keep aware of new techniques and possibilities under development. Demonstrations need to be publicized, too, and the success or lack thereof provides further information to both researchers and developers. Finally, the existence of the dissemination capability itself needs to be widely publicized to insure that all those having the need can make appropriate use of it.

Applied to any given piece of knowledge or any given improvement in one or a few schools, the R and D process and the implementation of improvement process appear to be identical. The differences between the two processes begin to become visible, however, when the systematic implementation of improvements in the entire educational system is the desired objective. The systematic and continuing introduction of improvements in education through judicious innovation and change based on research requires careful attention to the different roles played by the many independent agencies and institutions in education.

The first attempt, for example, to implement a new curriculum in economics in a secondary school undertaken with the aid of the original project staff responsible for developing it can easily be viewed as the normal demonstration responsibility of the project. But if the scope of responsibility is enlarged to include developing the capability of any teaching staff anywhere in the country where a decision to implement that curriculum has been made (that decision assumes, of course, knowledge about the curriculum and a judgement as to its suitability), the problem is clearly larger than an individual project. Attention needs to be paid to the problems of developing mechanisms for the inservice and preservice training of teachers to use the new curriculum. That effort in turn may point to the particular responsibilities of teacher training institutions across the nation both to assist in the particular effort and to insure, more generally, that preservice training is undertaken in the light of the almost certain need for continuing education of teaching staff to keep up with new departures in curricular application. Thus we have moved in this one instance very quickly away from the individual school setting to the much larger question of how teacher training institutions – and maybe, therefore, also state teacher certification agencies – are to play their particular and general roles of enhancing the likelihood of the adoption of proven innovation.

The matter can be carried still further. School boards should be apprised of the need to continually re-examine curricular objectives in terms of present and future social needs; as a consequence they can be expected to play a significant part in the networks of differing responsible agencies having unique roles in the process of implementing educational improvements based on and validated by research. The responsibilities of industry in helping to develop materials, in manufacturing, and in selling them similarly need careful attention. Their responsibilities, if any, for training their customers in the use of newly developed materials also need to be considered.

Viewed on a larger scale, then, it would appear that many of the questions associated with the actual implementation of solid innovations system-wide depend on factors more closely related to the dynamics of social change than to the actual research and development process itself. This dynamic — whether called administration, the politics of change, or continuing attention to the intricacies of the process of building conviction in and a demand for more effective materials and practices in the schools — is a necessary condition for the implementation of the products of R and D in schools. It constitutes, therefore, an absolutely essential problem for educational R and D strategists.

RESEARCH DEVELOPMENT AND DISSEMINATION STRATEGIES OF THE PAST

In describing the strategies of the past, present, and future we have elected to characterize broadly in generalized form the major outlines of attack as pursued until a very few months ago, the kinds of changes which took place and that have placed us in a crucial period of transition, and the sorts of directions we project for the future.

The availability of millions of dollars for educational research and development is a relatively new phenomenon. The recency of this growth and its absolute size give some pause for thought. A review of developments will help disclose the range of options available to us now and in the future.

The first piece of legislation specifically authorizing the United States Office of Education to support research, surveys, and demonstrations in education — The Cooperative Research Act — was passed by the Eighty-third Congress in 1954. In effect, the Congress said that research in education was important enough that the nation needed legislation to authorize Federal support for it. The Act itself, of course, represented a first step in the development of strategies for conducting R and D to improve education. Four years later, in 1958, the Congress included two titles in the National Defense Education Act (NDEA) authorizing research on the uses of new media and on modern foreign languages. These three authorizations turned out to be but a beginning. There are now a total of seven pieces of legislation including authorizations pertaining to vocational education, handicapped children and youth, libraries and information science, and an authorization under P. L. 480, a special foreign currency program based on the distribution of surplus foods.

The passage and funding of Federal legislation authorizing research in education was a major departure in and of itself. It was recognition of the conviction that Federal support needed to be directed to a problem bearing dimensions larger than those of individual states and calling upon resources, both financial and human, then currently beyond the means of state and local educational agencies.

The initial forays in support of educational research combined the identification of areas of concern and the support of technically excellent proposals submitted in response to a general call. In the very first instance, for example, proposals were requested dealing with mental retardation. As the research programs expanded in dollar size, greater proportions were administered on a nonpriority basis. Areas of importance were identified periodically and interest maintained in those areas, but the guiding principle for the support of research was the technical excellence of proposals rather than the type of research or the substantive area of concern. That principle was viable because in the early days of the research effort it was always possible to fund all the technically excellent proposals without regard to the field of study.

Over the years, substantial amounts of applied research as well as basic research were stimulated and supported. In more recent years the research program has moved into two kinds of activity: curriculum improvement and the establishment of research centers designed to explore programmatically R and D in a defined problem area in education. Even with the addition of these new thrusts the same principle mentioned before was still applicable. The field was young and growing, but it was still possible to support all of the quality proposals submitted.

THE RECENT PAST AND THE PRESENT

In 1965, several groups including staff inside the Office of Education, a Presidential Task Force, and others began to take a hard look at federally supported research in education, then nearly 9 years old. The wisdom of earlier strategies was apparent. They had been successful — in part because of the availability of dollars, in part because of the generally unsolicited mode of operation — in stimulating the growth of and interest in research and development. The field was growing; the submissions of proposals was increasing rapidly (in fact, at a faster rate than appropriations); the educational community as a whole was beginning to pay attention to the possibilities of research.

IMPORTANT RESEARCH DEVELOPMENTS

Research and Development Centers. The newly acquired visibility of research and development, however, led to concern over some of the apparent deficiencies of the total research effort as it had emerged to that point. First, it became apparent that the individual research projects as a whole — while the acceptable levels of quality individually — did not fit together well enough to be considered coordinated approaches to substantive problems in education. The Research and Development Centers were in part created in response to this need, but it was felt that major chunks of the research effort should be better coordinated and designed to lead to cumulative results.

Second, part of the difficulty in developing highly coordinated cumulative research efforts could be attributed to the inadequacy of the dissemination of information to the educational research community regarding the findings of completed research and the nature of current research. A strong need was felt, therefore, for the establishment of an effective research dissemination system.

Third, a careful examination of the outcomes of research in terms of service to the schools revealed that in two senses insufficient attention was being paid to the gap between the research stage and implementation. First, the stage of development was not being supported to anything near the degree that it should. Second, very little attention was being directed to the change process whereby improvements could be actually implemented in widespread fashion.

Fourth, it was clear that the human resources available for R and D activities would need to be expanded in at least two ways. Training programs to develop new talent would be required. New agencies and institutions previously not eligible for support would need to be tapped.

Amendments to Research Acts. The response to these felt needs took the form of a series of amendments to research legislation, the creation of a series of new program efforts under the aegis of the newly acquired authority, and the passage of a new piece of legislation. The basic piece of research legislation administered by the Office of Education, The Cooperative Research Act, was amended in 1965 by Title IV of the Elementary and Secondary Education Act (ESEA). The amendments broadened the existing authority – to support research, surveys, and demonstrations in education – to include dissemination. The range of eligible institutions was expanded to virtually all kinds of public and private organizations whether profit or non-profit. Authority was included to develop programs designed to train educational research and related personnel, and to upgrade training programs. The Office was authorized to award grants as well as contracts, and the Commissioner was given authority to award funds for the construction and equipping of facilities for research and related purposes.

These amendments vastly extended the range of activities possible under the research program, and made it feasible for the research program to meet directly some of the needs identified in the reviews of the program to that time. In addition, the testimony before the Congressional committees – or the legislative history, as it is called – made it very clear that the broadened authority was to be used to bridge the gap between research and practice and to pay substantially more attention to the problems of implementing the knowledge derived from the research efforts to date and in the future. The broadened responsibilities represented by these additional authorizations have created new demands on the administration of research and at the same time have offered new tools for meeting emergent shortcomings of the research effort.

Training Program. To date, three major new program thrusts have undergone spirited development. A training program designed to expand the corps of educational researchers has been developed. This program provides both program development grants and an array of training mechanisms ranging from institutes of short and long duration through undergraduate and graduate training programs to a small postdoctoral effort (Clark, 1967).

Educational Research Information Center. The Educational Research Information Center (ERIC) has been established with Central ERIC, twelve clearinghouses, and more to come. ERIC is designed to provide to researchers and practitioners alike an information storage and retrieval mechanism to make available instantly in easily accessible indexed form all the research and related data relevant to a particular problem. Research in Education, the monthly publication of the ERIC system, has now published several issues and already the usefulness of this effort has begun to prove itself.

Educational Laboratories. The third and largest development in the research program has awakened the interest, excitement, and the enthusiasm of the entire educational community from local schools, state agencies, and teacher training institutions to industry, scholars from the arts and sciences, and the lay public. We refer to the National Program of Educational Laboratories. Drawing from resources in regions extending across the nation, new institutions called educational laboratories have been created to bridge the gap between research and practice. These institutions were created by representatives of all the many agencies and institutions which play different roles in the implementation process. They are reflections of the conviction that it is not enough to do research; that research must be followed up by development project which, having established the desired objectives – whether curricular, instructional, organizational, professional, or technical – then move to the development of solutions drawing upon the best that research has to offer. The laboratories have also been charged with the responsibility for active dissemination campaigns based on the successful development projects they and others engage in.

We have encouraged the laboratories to conceive their responsibilities broadly – encompassing much more than merely the distribution of information, although that is a critically important function among many. Clearly, one of the important steps in the diffusion of research-based improvement throughout the educational system is the establishment of demonstrations. First time demonstrations of feasibility will be the direct responsibility of the laboratories; the more widespread diffusion of the successful innovations will depend on the degree to which information about the innovation is distributed to various parts of the systems, the degree to which the innovation recommends itself to professionals, and the degree to which credible demonstrations of the new practice or curriculum are mounted in schools. This last part of the diffusion process – the establishment of real-life (not hothouse) demonstrations of the innovation without the intervention of the inventor – represents a place where the labs can work with the state education agencies and local school districts to make use of Title III of ESEA. Such demonstration efforts must be coordinated and well-conceived. Laboratories can help only in their areas of competence and program thrust and only by cooperating closely with state education agencies and the local school districts.

The laboratories were conceived in such a way as to involve in their government and their operation the many types of different responsibilities and resources that combine to form the educational system as we know it. The hope is that these new institutions – responsible for their own program development and implementation and knit closely together to form a network – will, as their resources permit, at one and the same time engage in 1) major efforts to develop new materials, practices and organizations using the outcomes of research; and 2) by utilizing the expertise brought to them by the involvement of different agencies and institutions, pursue courses of action which help to hasten the process of improvement once tested innovations are available.

The strategy for these three new program thrusts is straightforward. If there is to be an expansion in research and development, people will need to be trained or recruited from new areas to fill the demand. In order to build structures of knowledge and cumulative improvement, dissemination networks and the material which moves through them will need to be better systematized and improved. Finally, to fill the gap between research and practice, a new autonomous institution drawing institutionally and representationally on many resources in the education system has been created.

TITLE III OF ESEA

In addition to the new authorizations for research and the programs mounted as a result of them, the Eighty-ninth Congress passed Title III of the ESEA. This program, authorizing support for projects submitted by local educational agencies designed to supplement existing school programs or to serve as models for existing school programs, was an innovation in itself. When (former) Commissioner Keppel testified before the House Subcommittee on the Elementary and Secondary bill, he told the Congressmen that he viewed Title III and Title IV together. Title IV was to be the means whereby the substance of educational improvement would receive increased impetus and attention; Title III would be the means by which local schools could inaugurate the kind of credible real-life demonstrations which would be convincing to their counterparts and become thereby one of the moving forces for the widespread adoption of tested innovation. Those responsible for Title III estimate that this past fiscal year fully 60 percent of the money went to support projects relating to the functions of development and demonstration. This year the program expects to do at least as well. Most important, for the first time local educational agencies were provided with the funds and the encouragement to experiment with new ideas.

CURRENT DEMANDS AND STRATEGIES FOR THE FUTURE

A thorough consideration of the present status of educational research on the basis of five variables provides important clues to the directions future strategies may well take. Those five variables are discussed briefly in the following paragraphs.

VARIABLES RELATED TO STRATEGIES

The Research Function. The research function has been charged broadly with 1) the basic research responsibilities relating to improving our understanding of the nature of learning and the conditions for it, 2) developing materials, practices, processes, and institutions designed to improve instructional practices, and 3) engaging in a range of activities running from demonstration to dissemination to training, in order to ensure that improvements find their way to implementation in operational settings.

The Financial Demand. The financial demand, represented by the broad responsibilities indicated above, is difficult to estimate with precision, but its dimensions can be sketched. Recent experience accumulated by the National Science Foundation, the Office of Education, and the Office of Economic Opportunity has shown that full-scale multi-media curriculum development can range in cost from $2 million to $6 million per course; even this expense is engendered by building curriculums for schools as they are presently structured. The costs for similar development for newly organized institutions might well be more. In any case, however, it does not take very complicated arithmetic to multiply the above figure by the number of possible course experiences that might be developed for preschool, elementary, secondary, community college, undergraduate, graduate, and adult education. If one assumes, as we think is reasonable, that such curricula probably ought to be redeveloped every 5 or 7 years it is not unreasonable to project an annual expenditure of a half billion dollars just for curriculum, hardware, and organizational development alone. The research effort to support that kind of development effort will add another $150 million to $200 million to that figure annually. The cost of demonstrations must also be added – perhaps as much as $300 million or $400 million annually. A rough rule of thumb for dissemination activities indicates that approximately 5 to 10 percent of the research effort ought to be spent in that area. We can add conservatively for that purpose another $50 million. The total is slightly more than a billion dollars annually, an amount which – if added to the current estimated expenditures of educational institutions for this fiscal year – would constitute just 2.2 percent of the total. That rate of expenditure for research and development approximates the minimum percentage devoted to R and D today by American industry. It strikes us as reasonable that the education industry, with estimated expenditures of $48.8 billion in fiscal 1967, should devote a similar proportion of effort to research and development in areas relevant to its missions.

The Dollars Available. The actual dollars available for the broad range of activities associated with the research effort have grown dramatically. They are by no means equivalent to the demands briefly sketched above. But major funding for research and related activities is still in its infancy. Significant educational research funding began only in 1957 at a combined USOE/NSF level of $1.7 million and has grown this year to an annual level of $200 to $220 million including USOE research appropriations, the NSF Course Content Improvement Activities, other miscellaneous agency expenditures, and Title III of ESEA (assuming that the 1966 proportion – 60 percent – of the Title III, ESEA, activities going to development and demonstration holds for 1967).[3]

This level of funding, matched with the estimate of the conceivable demand for funds for research and related activities, underscores the critical nature of the allocation of financial resources.

Manpower Resources. The problem of manpower resources is of crucial importance and is not easily resolved. A direct response, although a long-range one, is to train a new corps of professionals having the kinds of skills required to perform the entire range of research functions. There are other possible approaches, however. There already are – in other disciplines and in other kinds of institutions – individuals skilled in the research and development environment, and who could probably be persuaded to turn their attention to the field of education. These considerations were in no small measure part of the strategy behind broadening the list of eligible organizations to participate in R and D. They also played a major role in the early decision in the administration of research to secure the active involvement and participation of arts and sciences disciplines in educational research, a strategy that has continued until now and will be maintained in the future.

The combination of approaches adopted for the provision of adequate manpower supplies can be expected, then, to have considerable bearing on the possibilities for growth (see Clark, 1965; Guba and Elam, 1965; Hopkins, 1967).

Successses of Past and Current Efforts. The successes of educational research and development, the growing visibility of operations in this area, and the increasing conviction that research will in the long run provide the basis for substantive improvements in education have created new demands upon the human and financial resources available to the research enterprise. As soon as the demand outstripped the resources, something that happened only very recently, the entire picture of support necessarily was altered. It has now become necessary to allocate limited resources in the face of demands which exceed those resources.

ADDITIONAL STRATEGIES ADOPTED

The five considerations discussed above have led to the adoption of additional strategies. Broadly speaking those strategies embrace the three principal functions discussed below.

Careful Long-Range Planning. Now that the demand has clearly out-stripped the supply of resources, particularly in the financial area, it is critically important that the allocation of research dollars to various research functions and to particular topics or areas of development be accomplished in a careful and logical manner. This must be done in ways to ensure that both short-term and long-term interests are met, and to maintain and strengthen the existing political structure of education as it has developed over the years.

The careful long-range planning of those resources is much more likely to produce optimum levels of support for all the different functions and areas for research and development than is a totally non-directive approach embracing the competitive selection of technically excellent but unsolicited proposals for support. Such planning will need to take into consideration the instructional and administrative needs of the different educational levels and the many kinds of agencies, institutions and organizations in the field of education, industry, the professions, and government. Attention will have to be given to the needs of different target populations, including the general student, the handicapped, the disadvantaged, and the gifted. This implies, of course, that it will be necessary for us to develop much more effective mechanisms for planning research and related activities in order to ensure that all relevant data are canvassed and that maximum flexibility is built into the planning and administrative policies devolving from that planning. We must be careful, for example, that we continue to pursue the wisdom of earlier efforts in support of research by maintaining a significant portion of the research budget for unsolicited efforts. This is only one element of the larger planning picture, however, although a very important one.

As part of the attempt to improve the data on the basis of which research and other educational planning is now done, the Bureau of Research has recently called for proposals to begin pilot projects we hope will ultimately lead to the establishment of educational policy centers. These centers would, on a continuing basis, conduct the kinds of studies which would enable all of us to have better ideas, of the 5-, 10-, 20-, and 30-year picture for education. How should or might the schools relate to the society? What should the range of curricular objectives be? What kinds of sources might exist for financing education? How might schooling be structured in the society in the years to come? What technologies are likely to be available? These centers will not be engaged in predicting what the future will be. Instead, they will try to project the many alternatives available to us, cost them out, examine the consequences, and thus provide much better data to policy planners than are now available. They will provide a new kind of information and analysis looking to the future in a systems-oriented interdisciplinary way, supplementing the improved present-oriented data collection which must also accompany the planning process (see Goodlad, 1967).

The Outputs of Research Process. Our concern for the outputs of the research process is pragmatic in part, but it is tied closely to our conception of the mission of the research effort in education in general and of USOE's research program in particular. We have stated our faith that the improvement of education depends in the long run on the systematic prosecution of a well-conceived R and D effort. The improvement of education is the fundamental thrust of the mission; that criterion must be the continuing evaluative guide we employ in judging the success of our efforts.

While it is true that the justification for supporting research and related activities in education with public money can only be the eventual and significant betterment of the educational system as a whole, it is also true, happily, that the more we can demonstrate that research has affected school practices in a positive way, the more likely it will be that greater support for the entire research process will be forthcoming. Thus we believe that such objectives are not only good public policy, but good politics as well.

There can be no denying the growth and the strength of the interest in results. The Congress is concerned. The Executive Branch is concerned. Local school people have made their interest felt in a number of ways, particularly in the development of the programs for the educational laboratories.

We believe, however, that a strong and continuing tension exists between the conduct of research designed to increase our knowledge of basic learning processes, for example, and the pursuit of development projects designed to yield specific outcomes for instructional use. The demand for results has raised the appropriate and throughly justifiable concern of the academic and scholarly research community that such demands may tend to compromise the long-range efforts which are so badly needed. The continuing tension between short-term and long-term requirements – between today's youngsters and tomorrow's – is real and constitutes one of the continuing nightmares of the research administrator. This problem is the responsibility of the administrator of research, and it must be squarely met. We believe that it is important that the two functions not be confused, but that they both be served with sophistication and energy.

The importance of concentrating on the improvement of schools, and the significance of development projects in that effort, signals an increasing attention to such activities as curriculum development and explicit attempts to improve instructional practices. We are increasingly persuaded, for example, that prudent management of the limited resources available at the present time makes it necessary for us to adopt a research strategy that relates applied research projects closely to identified development efforts. We would hasten to add that the preservation of a signficant portion of funds for unsolicited proposals will still permit the support of some applied research unrelated to development since good ideas will also be generated independently of development projects or research planning groups and should be supported on their own merits.

We anticipate, then, still further increase in attention to development kinds of tasks designed to produce substantial, measurable, and cumulative improvements in the nation's schools.

Continued Expansion of Research and Related Activities. There is no question in our minds, as all the preceeding analysis should clearly indicate, that research and related activities must continue to expand in rapid but judicious ways. By directing our attention to the development of coordinated strategies for the support of research with the implication that greater attention will be paid to the outcomes of the research effort, we believe that constituencies in support of educational research will grow and the dollars available for the function will increase to what we believe are the regular minimal levels we suggested earlier.

SUMMARY

In this paper we have sketched out the two parts of the change process: research and development, on the one hand, and implementation, on the other. We have suggested that the principal purpose of the research effort ought to be viewed as the improvement of education in the schools. We have indicated the national need for research, and the importance of drawing on national resources to meet that need. We have described the discrepancy between the potential size of the need in dollar terms and the actual fiscal resources available. We have suggested that these conditions lead us to adopt, in effect, multiple strategies for R and D. These include: 1) better planning of the research effort; 2) more attention to the training of research and related personnel; 3) the establishment of new institutions – the laboratories – and new programs – Title III – to greatly increase the opportunities and responsibilities of local school officials in the research and implementation process; 4) the continuation of significant proportions of unsolicited research and development funds; and 5) increased attention to development efforts designed to yield better materials, equipment, and instructional practices for the schools. All of these strategies will, we hope, lead to increased understanding of the ultimate impact of the research effort, help to bring about substantial increases in the amount of support committed to research and development in education, and will provide the foundation for the continuing improvement and self-renewal of our educational system.

NOTE: Dr. R. Louis Bright is no longer with the United States Office of Education. He is presently located at Baylor University in Waco, Texas where he is a professor.

FOOTNOTES

1. Digest of Educational Statistics, 1966, U. S. Government Printing Office, Washington, D. C., 1966, p. 17.

2. The imprecision in the estimate derives from a number of considerations. The research authorizations of USOE in FY 1967 total $99 million, but $6 million are for training and another $14 million are for construction. The National Science Foundation spends close to $17 million a year on curriculum development, and almost certainly some of the curriculum development efforts of the Office of Economic Opportunity will be used by schools. Similarly, the National Institute of Mental Health and National Institute of Child Health and Human Development also spend a considerable amount on research related to areas of interest to learning and education. Title III of ESEA, if past percentages hold, could well spend $81 million on development and demonstration activities. Part of the development of strategies for educational R and D is the coordination of these several sources of research and research related funds.

3. The actual figures by program and year are as follows: (See chart on following page.)

	USOE Research	NSF Course Content	Title III, ESEA Dev. and Dem.
1954	$.....	$ 1,725	$.....
1955	.....	35,000	.....
1956	.....	18,000	.....
1957	1,020,000	650,140	.....
1958	2,300,000	750,310	.....
1959	6,716,000	6,180,485	.....
1960	10,350,000	6,302,055	.....
1961	10,117,750	6,167,740	.....
1962	11,770,000	9,389,948	.....
1963	14,188,400	12,626,771	.....
1964	19,820,000	14,157,650	.....
1965	37,703,000	14,889,081	.....
1966	100,141,241	16,393,383	45,400,000
1967	99,600,000	17,000,000 (est.)	81,000,000 (est.)

REFERENCES

Clark, D. L. Educational research and development: The next decade. In E. L. Morphet and C. O. Ryan (Editors), *Implications for education of prospective changes in society*. Reports Prepared for the Second AREA Conference. Denver, Colorado: Bradford-Robinson, 1967. Pp. 156-175.

Goodlad, J. I. The educational program to 1980 and beyond. In E. L. Morphet and C. O. Ryan (Editors), *Implications for education of prospective changes in society*. Reports prepared for the Second AREA Conference. Denver, Colorado: Bradford-Robinson, 1967. Pp. 47-60.

Guba, E., and Elam, S. (Editors) *The training and nurture of educational researchers*. Sixth Annual Phi Delta Kappa Symposium on Educational Research. Bloomington, Indiana: Phi Delta Kappa, 1965.

Hopkins, J. E. Scope of the demand for educational research and research related persons. Paper presented at the meeting of the American Educational Research Association, New York, February 1967.

The Wisconsin Research and Development Center for Cognitive Learning

Herbert J. Klausmeier
The University of Wisconsin
Madison, Wisconsin

THE AGREEMENT between The University of Wisconsin and the United States Office of Education (USOE) establishing the Wisconsin Research and Development Center for Cognitive Learning was signed on August 6, 1964. The high expectations held for the Center were expressed by the principals to the signing in a number of ways.

University President Fred Harvey Harrington remarked that the signing denoted "... an important day in the history of the University. This agreement marks the culmination of years of effort devoted to the improvement of the research capabilities of the School of Education, and opens the way to still greater developments."

Howard F. Hjelm, then director of basic research for USOE, described the agreement as following the new pattern in educational research of full institutional commitment. He stated: "The University has pledged itself to fully investigate the problem of learning and to disseminate research findings in a way which will bring about definite changes in school practices."

Hjelm found the University well suited for the commitment it had accepted. "The University was chosen because of the exceptional quality of its educational researchers, because of its excellent relationship with local schools and with the State Department of Public Instruction, and because of the wholehearted dedication to the project existing in Madison."

Center Director Herbert J. Klausmeier stated the primary focus of the Center: "Research emphases will be on learning by children and youth in normal school situations. Especially stressed will be the learning of concepts and problem-solving techniques in mathematics and other basic subjects."

The Wisconsin R and D Center was established under provisions of the Cooperative Research Program of USOE. In 1964 the Cooperative Research Program included six major areas: 1) basic and applied research, 2) demonstration, 3) curriculum improvement, 4) small contract, 5) research and development centers, and 6) developmental activities. Further description of the research and development centers was given in OE Publication 12017, 1963. as follows:

> Research and development centers are designed to concentrate human and financial resources on a particular problem area in education over an extended period of time in an attempt to make a significant contribution toward an understanding of, and an improvement of educational practice in, the problem area. More specifically, the personnel of a center will:
>
> 1. Conduct basic and applied research studies, both of the laboratory and field type.
> 2. Conduct development activities designed to translate systematically research findings into educational materials or procedures, and field test the developed products.
> 3. Demonstrate and disseminate information about the new programs or procedures which emerge from the research and development efforts. These activities may

include demonstrations in a natural, or operational, setting; the preparation of films, tapes, displays, publications, and lectures; and the participation in symposia and conferences.

4. Provide nationwide leadership in the chosen problem area.

The preceding ideas about the programmatic research, development, and dissemination thrust of an R and D Center in relation to a significant problem area in education followed logically from experience with other programs of the Cooperative Research Program of the Office of Education, the first of which, basic and applied research, started on a modest scale in 1956. Also, the research and development strategy had demonstrated a high yield in selected aspects of industry, agriculture, medicine, and space exploration. A large segment of the American public desired improved educational practices. Relevant USOE, university, and other concerned persons throughout the nation felt that a program of research and development centers would be equally effective in improving education.

An R and D Center, designed to improve educational practices through programmatic research and development, might be considered as a concept. As we know, concepts are abstractions of realities. A concept is comprised of the attributes, or properties, of things and events, the attributes being joined by rules, or ways of combining attributes systematically. An R and D Center then may be understood in relation to various attributes and how they are combined to form a whole. The writer hypothesizes eight crucial attributes of an R and D Center as follows: problem area, outcomes sought, research and development program and strategy, dissemination and adoption program and strategy, staff, administration and management, locus of operation, and source and stability of funding. The preceding attributes may be considered as the major components of a system, or instrumentality, designed to achieve certain general objectives implied by the problem area. Thus, a systems approach is properly followed in organizing and managing the Center operations and programs. The remainder of this chapter is devoted to a discussion of the various components of the system, their relationships, and the outcomes attained thus far by the Center staff.

Before further description of the Wisconsin R and D Center, it should be recalled that the Elementary and Secondary Education Act of 1965 as noted by Bright and Gideonse in the previous chapter, provided for the establishment of regional educational laboratories and Title III demonstration and adoption centers in local schools. These two programs of USOE are much larger than the current R and D Center Program. As instrumentalities for improving educational practices, the writer hypothesizes that they, too, can be described and understood in terms of the preceding eight attributes of a center. It is not the purpose of this chapter, however, to point to the similarities and differences among the various research, development, and demonstration instrumentalities of the USOE but to describe the R and D Center of Wisconsin as an example of a significant contributor to the extension of knowledge about cognitive learning and the improvement of related educational practices through research and development.

PROBLEM AREA

The problem area of the Center is the improvement of education through a better understanding of cognitive learning. The importance of this problem area may be inferred from five propositions regarding learning and instruction.

The first proposition is that concepts provide much of the basic material for thinking. Concepts possessed by the individual enable him to interpret the physical and social world and to make appropriate responses. Concepts once learned then serve as symbolic mediators between sensory input and overt behaviors. Without concepts with which to think, man, like lower-form animals, would be limited mainly to dealing with sensorimotor and perceptual representations of reality, closely tied to immediate sensory experiences.

A second proposition is that concepts and cognitive skills comprise the major outcomes of learning in various subject fields. Sentence, noun group, and transforms represent four concepts in English syntax. Set, number, and fraction represent concepts in mathematics. In order to learn the concepts of a discipline, a student must acquire cognitive skills. In mathematics these skills are designated by such terms as arithmetic comprehension and problem solving. The identification of concepts and their arrangement in a hierarchical order and the identification of cognitive skills are principal concerns of scholars and curriculum workers in the various subject fields.

A third proposition, related to the preceding ones, is that the acquisition of cognitive skills enables man not only to acquire the knowledge of others but also to generate new knowledge. Cognitive processes, such as cognizing, hypothesizing, and evaluating information, are learned. These and other processes, combined with cognitive contents, are designated cognitive abilities or skills. When the child has learned the cognitive skills of reading, he can acquire knowledge in the absence of other persons. When he has mastered mathematical concepts and computational skills, he can find solutions to problems not encountered in the same form by any other person.

A fourth proposition is that efficient learning of concepts and cognitive abilities is related to conditions within the learner and conditions within the situation. Two conditions within the learner associated with efficiency of learning are attention, related to motivation, and level of cognitive functioning. In school settings, the preceding conditions related to the student and other conditions related to the content and sequence of instruction, quality and availability of equipment and materials, and quality and methods of instruction are paramount. More efficient learning can be attained through better control of these conditions.

A final proposition is that knowledge about cognitive learning, concepts, cognitive skills, and conditions of learning generated by scholars requires validation in school settings. Although the study of cognitive learning is in its infancy, sufficient knowledge has been generated to develop instructional theory regarding the teaching and learning of concepts. The validity of the theories must be tested in school settings. Similarly, instructional materials for children and outlines of concepts for teachers must also be tested for usefulness and validity in school settings.

PROGRAMS, STRATEGIES, AND OUTCOMES

Strategies best suited to achieve the Center goals include the differentiation of the problem area into research and development programs and other programs. The three research and development programs are titled Processes and Conditions of Learning, Processes and Programs of Instruction, and Facilitative Environments; the other programs deal with Dissemination, Training Activities, and Technical Support. These programmatic activities are designed to concentrate and utilize human and monetary resources most effectively over an extended period of time. Each research and development program is characterized by a sufficiently sharp focus to assure its identity and by a relatedness of content and methodology that facilitates the free flow of information and personnel among programs. Figure 1 shows the major operation of the Center and the program components. These operations and components are now described in some detail, with attention first directed to the three R and D programs and later to the other programs. The objectives of each program are given and outcomes attained through 1968 are summarized.

PROGRAM 1 -- CONDITIONS AND PROCESSES OF LEARNING

To increase our understanding of conditions and processes of cognitive learning, basic research is needed. To test in school settings the knowledge from basic research, relevant models and systems must be developed. To meet these requirements, the research and development activities comprising Program 1 are directed toward: a) generating new knowledge about concept learning and cognitive skills by conducting laboratory-type experiments or research within the school setting; b) synthesizing existing knowledge and developing general taxonomies, models, or theories of cognitive learning; and c) developing educational materials suggested by the prior activities.

Substantive information from the basic studies in learning feeds directly into the area of instructional improvement, thus providing from the behavioral sciences a foundation for the development of instructional programs. Also, methods developed in the behavioral sciences are utilized in the research and development activities of Program 2. At the same time, the behavioral scientist in Program 1, when dealing with the learning of content in a subject field, utilizes personnel from Program 2 and other sources who are expert in the subject field, familiar with instructional processes, or both.

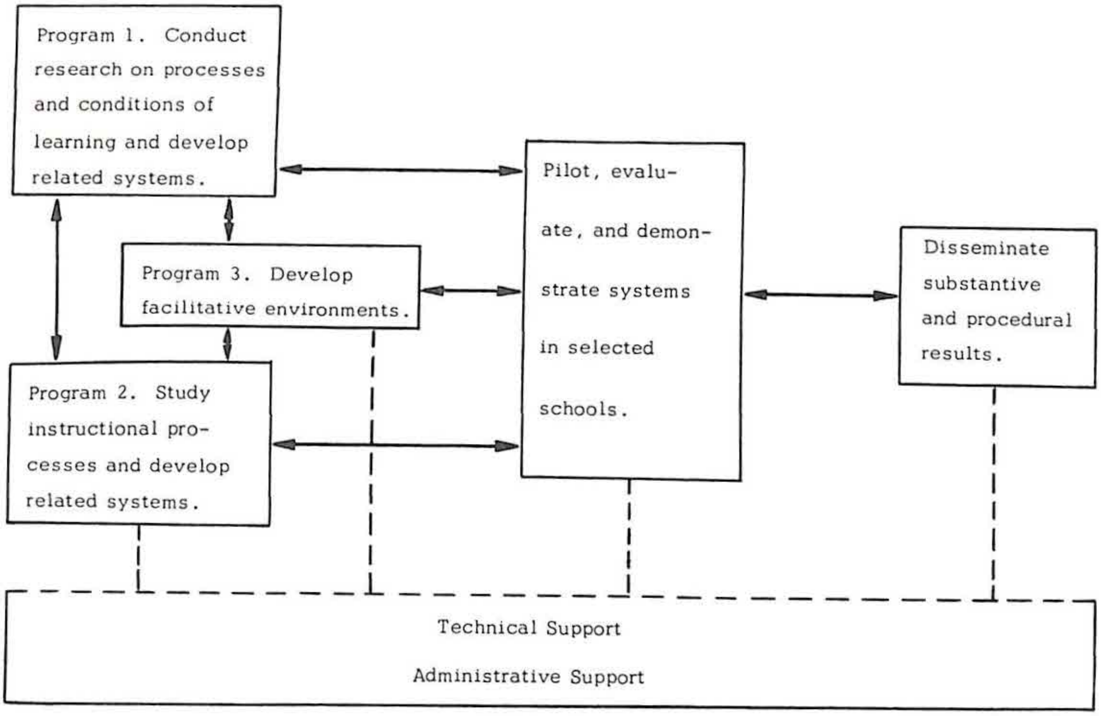
Basic research in cognitive learning has been conducted by the Center staff and reported regarding concept learning, creativity, prereading skills, rule learning, the shaping of complex human behavior, group pressures on individual performance, and the transmission properties of various media. Relevant results from these studies have gone into the development of instructional programs. In addition, a system of motivation was formulated and put into operation in selected elementary schools during 1967-68. Creativity training materials and procedures were prepared and received initial testing in selected schools in 1967-68. Procedures for teaching severely retarded readers were developed and field tested, starting in 1965, and were used operationally in selected schools in 1967-68. Other basic research dealing with concept learning and cognitive skills is continuing and is being applied to improvement of instructional programs in arithmetic, English, reading, and science.

PROGRAM 2 -- PROCESSES AND PROGRAMS OF INSTRUCTION

What is an appropriate unit to deal with when attempting to improve instruction, or stated differently, to achieve greater efficiency of student learning? The Center staff and school people decided that it is each building or an integrated set of buildings which the same children attend rather than individual children, separate classrooms within a building, or all the buildings of a local school system. Within each building, the instructional and other staff must deal with all the components of an instructional system. Similarly, the R and D staff, particularly those in Program 2, are concerned with all the components of an entire instructional system, although they may give much more time and effort to developing and testing one component rather than to another. Figure 2 outlines the major components of a total instructional system that are relevant to a school building. The order from top to bottom also indicates the general sequence in which the staff of the R and D Center and schools move in developing a prototypic system, such as in reading, mathematics, science, and English. Variables associated with each component require research. Thus, research on instructional processes and the development of instructional programs are essential in this program. The specific objectives of Program 2 are: a) to establish a rationale and strategy for developing instructional systems in the cognitive domain; b) to identify by careful synthesis and further research sequences of concepts and cognitive skills within and across disciplines; c) to develop assessment procedures and materials for the concepts and skills identified above; d) to identify existing materials or develop new instructional materials associated with the concepts and cognitive skills; and e) to generate new knowledge through research about instructional procedures including motivation, individualization, classroom management, and organization of instruction.

Executing the activities of this program requires specialists in a subject field, in instructional methods, or both, with assistance from behavioral scientists and communication experts. Each project in this program has at least one staff member whose speciality is cognitive learning. In addition, the program utilizes the many resources outside the Center including the University educational TV station (WHA-TV) and Educational Testing Service.

Figure 1: Major operations and components of the Wisconsin R and D Center for Cognitive Learning


Program 1. Conduct research on processes and conditions of learning and develop related systems.
Program 3. Develop facilitative environments.
Program 2. Study instructional processes and develop related systems.
Pilot, evaluate, and demonstrate systems in selected schools.
Disseminate substantive and procedural results.
Technical Support
Administrative Support

The largest instructional program developed by the Center staff through 1968 was Patterns in Arithmetic, a program comprised of TV video tapes, student lessons, and teacher notes. It is designed as a complete program of materials for children and also for simultaneous in-service education of teachers. As of 1967-68 the materials for Grades 1-4 had been developed and initially tested and were undergoing further field testing in many schools of Alabama and Wisconsin. The program for Grade 5 was being completed. The development of a prototypic instructional program, not using TV procedures and materials, dealing with concepts in mathematics, Kindergarten-Grade 6, was started in 1967-68.

Fundamental research was conducted and reported on the learning and teaching of reading, 1966-68, and is continuing. An outline of major content, behavioral objectives, and related appraisal procedures was developed, Kindergarten-Grade 6, and relevant instructional materials and procedures were identified. The entire prototypic instructional system was piloted in selected elementary schools, 1967-68, as cooperative school and Center activities, with substantial support of ESEA Title III and Title I funds.

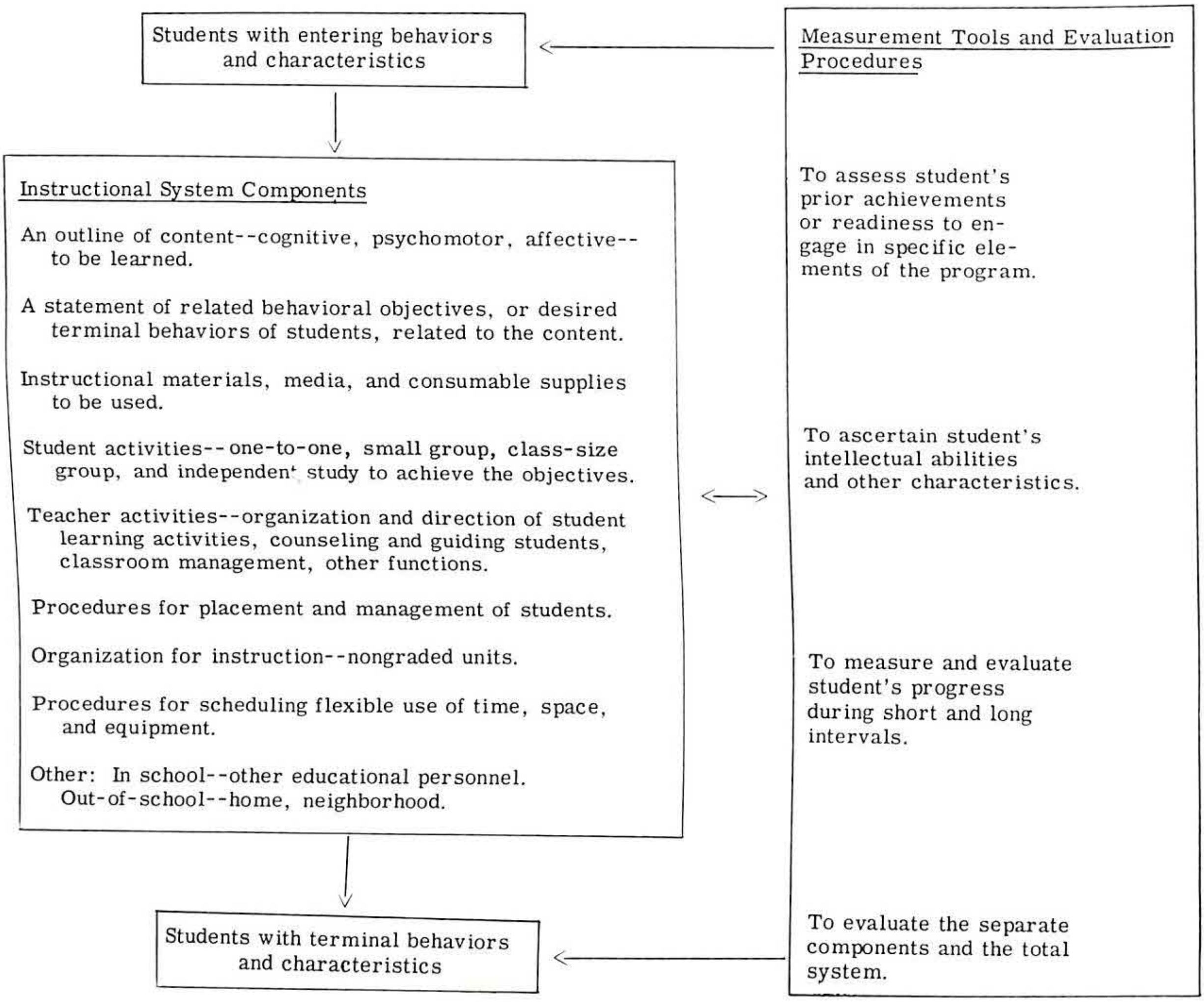
An outline of concepts related to the particle nature of matter and environmental conservation, appropriate for the elementary-school science program, was developed, 1965-68. Research on instructional methods was reported and other research is continuing. Preliminary instructional materials and methods appropriate to the concepts were developed during 1967-68 and initially tested. Related research and the development of scientific equipment and materials is continuing.

A prototypic program of instruction in English in the junior high school was under development, starting in 1966. An outline of selected concepts of English syntax and a related set of programmed instructional materials to teach the basic concepts were developed and field tested during 1966-67. Further testing of the programmed material, and implementation of the entire curriculum, was started in selected junior high schools during the 1967-68 school year and is continuing. Basic research on processes and conditions of learning is continuing.

An outline of concepts in verbal argument was developed, 1965-67, and related tests of critical thinking were also developed and initially validated. Programmed material was prepared during 1967-68 to teach these concepts. Further research on the program and validation of the tests continued, 1967-68.

Two other sets of activities have high priority for initiation in subsequent years. Research and related development regarding social studies in the elementary school are projected. Greatly increased use of the computer in research and in the management of individually guided education is the second field to be given increased emphasis.

Figure 2: Major components of an instructional system



Students with entering behaviors and characteristics

Instructional System Components

An outline of content--cognitive, psychomotor, affective--to be learned.

A statement of related behavioral objectives, or desired terminal behaviors of students, related to the content.

Instructional materials, media, and consumable supplies to be used.

Student activities--one-to-one, small group, class-size group, and independent study to achieve the objectives.

Teacher activities--organization and direction of student learning activities, counseling and guiding students, classroom management, other functions.

Procedures for placement and management of students.

Organization for instruction--nongraded units.

Procedures for scheduling flexible use of time, space, and equipment.

Other: In school--other educational personnel.
Out-of-school--home, neighborhood.

Students with terminal behaviors and characteristics

Measurement Tools and Evaluation Procedures

To assess student's prior achievements or readiness to engage in specific elements of the program.

To ascertain student's intellectual abilities and other characteristics.

To measure and evaluate student's progress during short and long intervals.

To evaluate the separate components and the total system.

PROGRAM 3 -- FACILITATIVE ENVIRONMENTS

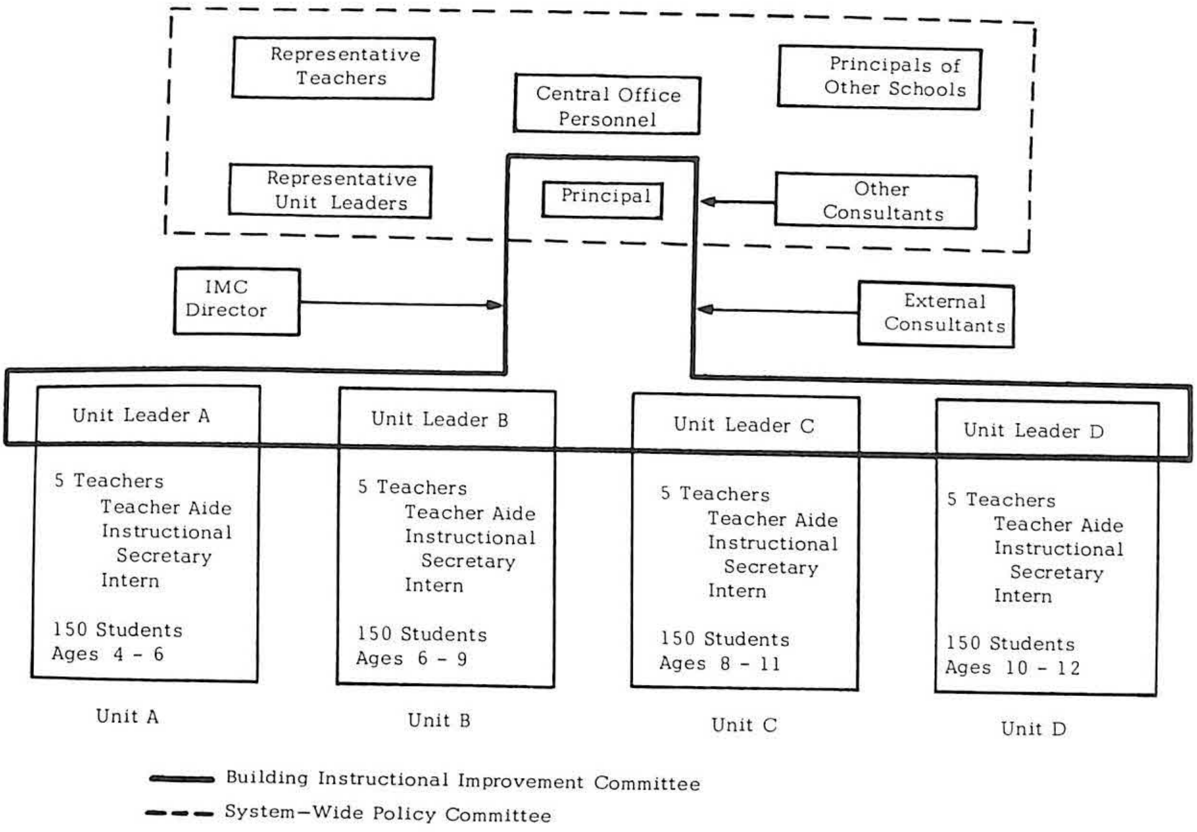
Achieving improved educational practices requires facilitative environments in selected local school buildings, in total school systems, and among various agencies. Thus, the research and development activities within this program are designed a) to develop and test new organizations and relationships among staff that facilitate research and development activities on cognitive learning in schools, and b) to develop and test the effectiveness of the means whereby schools select, introduce, and utilize the results of research and development. The same schools that facilitate student learning also serve as demonstration centers for entire school systems and the R and D Center.

In seven school systems of Wisconsin environments were established, 1965-68, designed to facilitate the research and development activities of the Center and the learning of children through a system of individually guided education. A new organization for instruction that replaces the self-contained classroom by instruction and research units and an instructional decision-making committee is an integral component of the facilitative environment of the Multiunit School. The organization is shown in Figure 3. The characteristics of individually guided education and of the organization which facilitates its execution merit further attention.

The Multiunit School organization includes both a formal organizational structure and a procedural style consisting of several essential components as shown in Figure 3. The organizational hierarchy of the Multiunit School consists of groups at three distinct levels of operation.

At the classroom level are the Instructional and Research (I and R) units. Each I and R unit has a unit leader or professional teacher, two or more regular staff teachers, one or more aides or secretaries, and in some cases an intern. The number

Figure 3: Organizational chart of a multiunit school of 600 students



of instructional staff varies with the size of the unit. The intern assumes instructional responsibilities and does not perform routine and clerical duties. Each unit is charged with planning and conducting the total school experience of about 150 students in the prototype shown in Figure 3.

At a second level of organization, the principal and the unit leaders constitute the permanent Instructional Improvement Committee of the building. The principal chairs the group, which meets weekly, more often if necessary. This committee may bring in consultants from the central office, the state education agency, or other agencies. Instructional decisions made by the committee are executed in the units.

At the third organizational level is the System-wide Policy Committee. Chaired by the superintendent or his designee, this committee includes representative principals, unit leaders, teachers, consultants, and other relevant central office staff. The personnel of this Committee have the capabilities and specialization necessary to assist the staff of the Multiunit School. The Committee meets less frequently than either of the other groups, but its operation is one key to the success of the Multiunit School.

The organizational pattern of the Multiunit School thus differs from that of the traditional, self-contained classroom school in several ways. First, in the Multiunit School personnel work in units or committees, rather than in isolation as is the case in the traditional school. Second, three new roles are added: unit leader, teacher aide, and instructional secretary. Finally, the addition of new roles and the use of personnel in groups rather than alone results in considerable redefinition of the familiar roles of principal, teacher, and consultant.

In individually guided education in the Multiunit School, the Instructional Improvement Committee determines the objectives for the particular school building, taking into account system-wide and state regulations. These are broad institutional objectives for the school building. The staff of each nongraded Instruction and Research Unit, the replacement for the graded homeroom or self-contained classroom, then decides the objectives for each child in the unit. While the unit leader takes the initiative here, each unit teacher also participates. Assessment of the child's characteristics is through observation and by means of locally constructed and standardized instruments of various types. On the basis of the assessment each child is then placed in one-to-one, small group, class-size group, and unit-size group activities. Instruction employing tutorial work and computer-assisted instruction are examples of one-to-one activities. Activities in small groups of two to fifteen are organized to attain socializing and also skill objectives. In connection with skill objectives, 150 children in a unit might be placed in fifteen small groups for most of their mathematics instruction and then regrouped in another fifteen groups for part of their reading instruction. Class-size or homeroom activities are used for achieving any objective where heterogeneity is desired. Large unit groups of 40 to 150 are formed mainly for giving information to the total group or for independent study. The information is given by a teacher, television, sound motion picture, or by other means. Some music and physical education activities are conducted in groups larger than the usual class size. Independent study is carried out in small groups, class-size groups, and large groups.

During 1967-68 approximately 4, 600 students in the elementary school, 1, 600 in the junior high school, and 400 in the senior high school were enrolled in R and I Units in Wisconsin schools. In these schools, children, educational practitioners, behavioral scientists, and other staff members of the Center are brought together in a close working relationship. Thus the selected schools of Wisconsin providing the facilitative environment comprise a primary link between improved school practices and Center research and development.

DISSEMINATION

The R and D Center is committed to disseminate the products of its research and development activities to appropriate personnel by various media and also to install them in selected schools where others may directly observe the related educational practices. So that the Center may fulfill its obligation of improving educational practice nationally, cooperation with personnel in key organizations of the state and nation is essential. An internal information section and an external network involving the Center and other agencies were established, starting in 1964, and are being extended in order a) to transmit information directly to scholars and others through Center supported publications, conferences, etc.; b) to communicate information to personnel of selected school systems and school-related educational agencies regarding the outcomes of the Center's programs as they become ready for dissemination and school use; c) to distribute knowledge regarding cognitive learning and products to agencies that function as linking mechanisms with schools; and d) to receive from the other agencies in the network information that may be used to improve the Center's products and operations.

Reports of research and descriptions of developmental activities are disseminated through various publications. In addition, all instructional programs and procedures produced and tested by the Center may be observed in selected schools. The Center is working with other agencies to assure statewide and national dissemination and adoption of Center products. The State Department of Public Instruction of Wisconsin, regional educational laboratories, Title I and Title III programs of ESEA, local schools of Wisconsin and other states, agencies of The University of Wisconsin, and the private sector are all involved in the dissemination network.

The Center's publication program carried out by the Office of Information Services has four major components – Technical Reports, Theoretical Papers, Working Papers, a Newsletter, and books. Technical reports describe the rationale, methods, and results of original research conducted by Center staff. Research reviews, statements of theoretical position, and descriptions and rationale for project development are the content of theoretical papers. The theoretical papers are formal documents that, like technical reports, may be classified as monographs. Practical papers contain information of immediate significance to educational practitioners in establishing or extending instructional programs – guidelines for beginning new programs, outlines and specific procedures for instruction, evaluation summaries and procedures, and progress reports of projects conducted in the schools. Working papers present plans and procedures for initiating and executing research and development activities.

Analyses of Concept Learning edited by Herbert J. Klausmeier and Chester W. Harris and the present book are collections of papers read and discussed at the Center. The first, from a conference of the same title, was published by Academic Press in the fall of 1966. The second, from the Center colloquium series, was published by Dembar Educational Research Services in 1968.

TRAINING ACTIVITIES

Three conditions indicate the need for a high priority to be given to the training of personnel in educational research and development in the R and D Center. First, the nation needs a large number of able persons in educational research and development. The many new programs under the ESEA of 1965 and others originating in the various states and local school systems require research and development personnel as well as knowledge and solutions to many educational problems. Second, The University of Wisconsin has both a tradition and the resources to contribute to the training of educational researchers, particularly beyond the baccalaureate. The University has strong graduate programs through which students acquire knowledge and skills essential for educational research. Third, the R and D Center at Wisconsin is the largest agency in the Midwest with educational research and development as its primary function.

The informal training of educational researchers and developers to supplement the formal programs in the various instructional departments and colleges, has been a significant contribution of the Center to improving education through research and development. For example, during 1966-67, informal training was provided for four postdoctoral fellows, sixty-three graduate students employed part time by the Center, ten ESEA Title IV doctoral fellows, ten NSF fellows, five NDEA fellows, several University fellows, and forty-two learning specialists employed by the public schools. Twenty-eight persons affiliated with the Center were awarded PhD's by the end of 1967 and are located throughout the country in a variety of leadership positions. A large group of persons has been awarded the Master's degree also.

TECHNICAL SUPPORT

The primary components of the Technical Section are a measurement and design group; production specialists; and computer experts. The design group provides assistance to the research personnel on matters of research design and evaluation. Equally important, this group is continuously developing new designs to meet the needs of the Center and refining statistical procedures for the analysis of data. A measurement expert provides assistance on problems of measurement and develops more efficient measurement tools and procedures. Computer specialists coordinate data analysis, from the conversion of raw data to computer compatible form through the presentation of the processed data in tables and graphs, and assist in the application of computer technology to educational research and development. A curriculum specialist assists the staff in identifying and using appropriate reference materials and in the production of graphical materials. Production experts assist with the development of instructional, experimental, and informative materials in the various media.

STAFF

Execution of the Center programs requires proper interdisciplinary staffing. This Center subscribes to the philosophy that the most productive research and the development of effective instructional systems will derive from teams composed primarily of behavioral scientists, subject-matter specialists, and experts in curriculum and other areas closely related to educational practice. The behavioral scientist contributes his research skills and knowledge of fundamental processes in cognitive learning; the subject-matter specialist provides knowledge of the subject field; the experts in curriculum and other fields contribute knowledge about methodology and other relevant instructional variables.

Studies of cognitive learning, included in Program 1, are headed by behavioral scientists concerned primarily with synthesizing and advancing knowledge in this area. The other experts contribute to these studies as team members when cognitive processes related to specific subject content are studied. In Program 2 in which processes of instruction are studied and instructional programs are developed, the projects are headed by subject-matter or curriculum specialists concerned with developing an instructional system in a subject area; however, a behavioral scientist is a member of the team. Experts in school organization and the sociology of groups are involved with other staff in the development and maintenance of the facilitative environments, Program 3. Communication specialists carry out the dissemination program.

This mutually beneficial involvement of the staff has two natural consequences. First, the basic studies in learning are concerned with developing knowledge that is relevant to educational practices. Second, the studies of instructional processes and the instructional products that are developed in Program 2 have a sound foundation in the behavioral sciences and the subject disciplines.

In addition to the Center staff, many outstanding interdisciplinary groups are available within the University for cooperative planning and execution of activities. Personnel from three of the outstanding activities in The University of Wisconsin are continuously engaged in consultation and team work with the Center staff: WHA-TV, a foremost producer of educational TV programs in the nation; the Survey Research Laboratory, a large and distinguished interdisciplinary survey team; and the Computing Center, one of the best in the nation. Thus, the Wisconsin R and D Center does not require a large permanent staff in each of these areas of specialty. Rather, programs and activities are planned and executed cooperatively as part of the total University of Wisconsin contribution to improvement of education through research and development. Other relevant interdisciplinary centers and laboratories at the University include the Center for Demography and Ecology, Mass Communications Research Center, and a Social System Research Institute.

Twenty-two persons of professorial rank were affiliated with the Center, 1967-68. The departments in which they held rank included Computer Science, Curriculum and Instruction, Educational Administration, Educational Policy Studies, Educational Psychology, English, Psychology, Sociology, and Speech. Some professors held rank in two departments.

ADMINISTRATION

Administratively, the Center is located in the University's School of Education but draws upon the entire University for human and other resources. The working relationship between the Center and the University is manifested in three ways: professors in various University departments are members of the Center's Faculty; both the Dean of Education and the Chancellor of the University are involved in reviewing the Center's activities; and University officials serve on the Center's Executive Committee, the Policy Review Board, and the Advisory Council. Figure 4 outlines the administration and management of the Center.

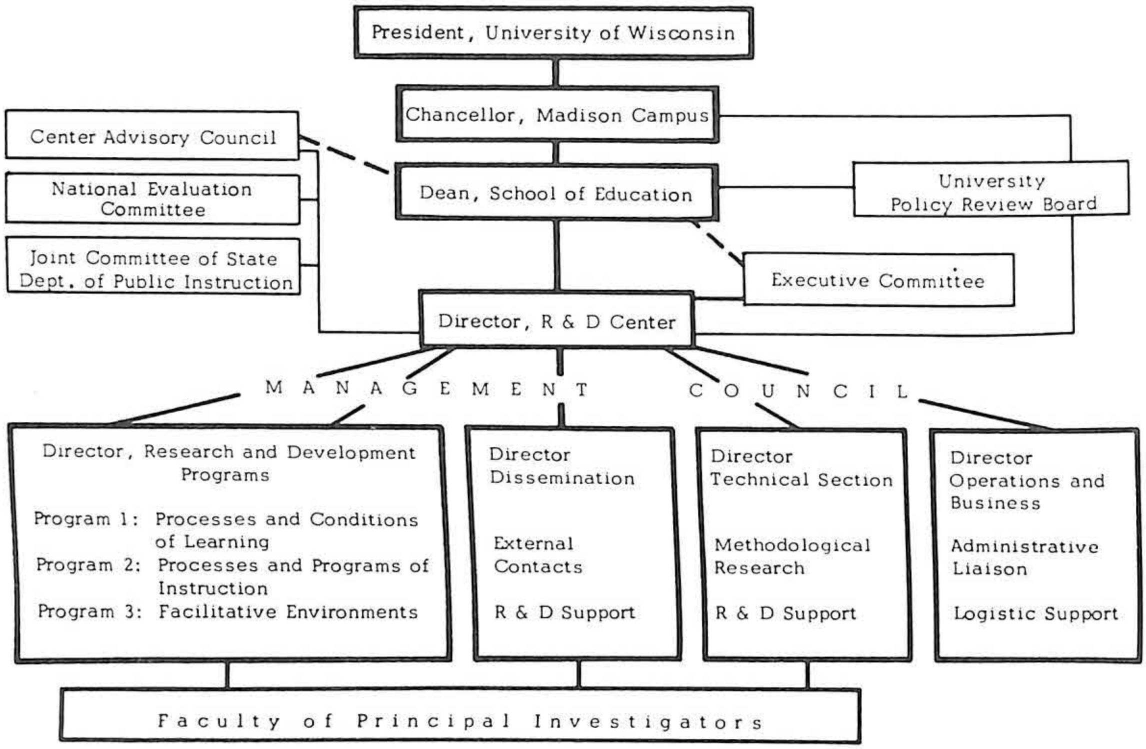
The Center Director is responsible for planning the total Center program and its broad strategies. A director of the research and development programs, a director of dissemination, a director of operations and business, and a director of a technical section, with the Center Director, comprise a Management Council.

The Executive Committee makes recommendations regarding policies, programs, and budget. It consists of the Center Director, who is also the chairman, three members elected by the Principal Investigators, and three members appointed by the Dean of the School of Education.

A Faculty of Principal Investigators, consisting of professors in various University departments, participate in the Center's program development, and they plan and execute the Center's projects.

The Policy Review Board brings the necessary University resources to bear upon the problems which the Center is interested in solving. The

Figure 4: Wisconsin Research and Development Center for Cognitive Learning organization chart


President, University of Wisconsin
Chancellor, Madison Campus
Center Advisory Council
National Evaluation Committee
Joint Committee of State Dept. of Public Instruction
Dean, School of Education
University Policy Review Board
Executive Committee
Director, R & D Center
MANAGEMENT COUNCIL
Director, Research and Development Programs
Program 1: Processes and Conditions of Learning
Program 2: Processes and Programs of Instruction
Program 3: Facilitative Environments
Director Dissemination
External Contacts
R & D Support
Director Technical Section
Methodological Research
R & D Support
Director Operations and Business
Administrative Liaison
Logistic Support
Faculty of Principal Investigators

Board is composed of representatives from the office of the Chancellor of the Madison Campus; the deans of the School of Education and the College of Letters and Science; the Director of the Wisconsin Improvement Program; the Superintendent of State Department of Public Instruction; a member of the Wisconsin Coordinating Committee for Higher Education; the Executive Secretary of the Wisconsin Education Association; three departmental chairmen of the College of Letters and Science and the School of Education; and the Center Director. The Board annually reviews the Center's mission, organization, programs and staffing; recommends overall policies; suggests funding levels.

A National Evaluation Committee confers annually with Center personnel on the Center's progress. The Committee consists of nine individuals of national reputation interested in the problem area of the Center and committed to the improvement of education through research and development activities.

The Center Advisory Council, composed of officers of the University and of a representative of the Wisconsin School Board Association, meets once a year to analyze the Center's progress and to advise on matters in need of attention. Individuals of sixteen school systems, including public and parochial schools and a technical institute, are included.

LOCUS OF OPERATION

Because the R and D Center is effectively integrated into the University complex, both on an administrative and disciplinary level, it is natural that the Center program incorporates the philosophy of The University of Wisconsin regarding the improvement of society through research, development, and training. This philosophy is best expressed in a widely quoted statement:

> Whatever may be the limitations which trammel inquiry elsewhere, we believe that the great state University of Wisconsin should ever encourage that continual and fearless sifting and winnowing by which alone the truth can be found. (Taken from a report of the Board of Regents in 1894.)

President Harrington has from the beginning implied the extension of these concepts to educational research through the R and D Center, and he continues to give generously of his time to assist the Center in achieving its goals, and to assure a climate in which the Center may operate more effectively.

The Madison Campus of The University of Wisconsin is the foremost research agency in the state and is among the leading research-oriented universities in the world. The commitment to educational research is substantial as may be inferred from a statement by then Chancellor Robben Fleming on July 1, 1965:

> With so much emphasis on education in today's world, it would be easy to drift along the well-marked paths of the past. The strength of our School of Education lies in its restless innovation. In cooperation with the schools of the state, it is forever trying new ways to improve the primary and secondary schools. In residence its curriculum is marked by an unusual degree of cross-fertilization, and by an interdisciplinary faculty. Its research is probing the frontiers of the learning process, asking, in effect, whether we are on the right path to begin with.

Implementation of the philosophy implied in the preceding statements as applied to the R and D Center has been demonstrated in two notable ways. First, the Policy Review Board of the Center, the most powerful governing agency of the R and D Center, is chaired by James Cleary, Vice-Chancellor of Academic Affairs. This Board reviews the program of the Center and takes major responsibility for assuring adequate staffing of the Center in light of its program. The second example of implementation is that the actual state support, not an indirect contribution, to the Center has increased from approximately $10,000 in 1965-66 to $108,000 in 1967-68. This state contribution is used almost totally to pay the salaries of principal investigators of the Center – professors who hold appointments in an academic department of the University and are budgeted fifty percent or more in the Center.

The School of Education is committed to the philosophy that the continuous improvement of public school education requires unceasing research and related activities. Dean McCarty, shortly after arrival at the University in 1966, indicated that the R and D concept is the most significant educational innovation in recent decades and that all forms of worthy educational research merit support. Three of many possible examples point to the high priorities given to educational research and related activities at The University of Wisconsin. First, $4,024,298 was expended on educational research and program experimentation in 1965 through the School of Education. The four sources of the support were Federal agencies, the State of Wisconsin, private organizations, and local school systems. Second, in the first seven years of the Cooperative Research Program of USOE, the University was awarded grants for thirty-three projects. Third, the amount of state support for research to the School of Education increased from $294,963 in 1964-65 to $804,976 in 1966-67.

The interdisciplinary nature of the School of Education is a major factor in the eminence of the University in educational research and related activities. In 1965 the membership in the School of Education was 1,075 professors from seventy-two departments of the University. Graduate work in preparing educational researchers is interdisciplinary. Professors in many departments of the University offer courses taken by graduate students pursuing programs to prepare for educational research. Further, many PhD candidates in the various departments of the School of Education pursue minors in departments of the College of Letters and Science while many PhD candidates in Letters and Science pursue minors in departments of the School of Education.

SOURCE AND STABILITY OF FUNDING

Any organization requires stable funding over a period of years in order to secure excellent staff at reasonable salary levels. Stability of funding was an important consideration in starting the Center. This Center, in line with the general USOE policy, was awarded a 5-year grant, starting September 1, 1964, by the USOE with three important provisions: First, that the contract would be amended annually; second, that, if for any reason, it became necessary to discontinue the Center the phasing out process would be accomplished over a period of time mutually satisfactory to The University of Wisconsin and USOE; and third, in the fourth year, the achievements and operations of the Center would be evaluated comprehensively in order to arrive at a decision regarding a second 5-year period. The Center was awarded approximately $500,000 for the first year of operation. Each amended contract for the next three years showed a modest increase in the level of support.

The second source of support is the State of Wisconsin through The University of Wisconsin. As noted earlier, some monies are made directly available to the Center. This is, however, not the main University support. The largest support lies in the availability of personnel to the Center – administrators, professors, graduate students, and classified personnel – and also other personnel in specialized fields such as computer science, opinion survey, TV production, and equipment development. The University also has leased a facility and provides the many services that are necessary to make an operation function well.

Although not a direct source of funds to support Center activities, the staff of four local school systems, particularly the subject-matter consultants, the building principals, and unit leaders, give considerable time to refining the concept of individually guided education in multiunit schools. Also, ideas and innovations of the local schools and the R and D Center are demonstrated through Title III projects in local schools. The State Department of Public Instruction is the principal disseminator of R and D field-tested ideas and materials in Wisconsin.

CONCLUDING STATEMENT

The preceding description of the Wisconsin Research and Development Center for Cognitive Learning was written late in the third year of the Center's existence. It is anticipated that there will not be sharp changes during the next decade in any of the eight major components: problem area, outcomes sought, research and development programs and strategies, dissemination program and strategies, pattern of staffing, pattern of administration and management, locus of operations, and source and stability of funding. Though sharp changes are not expected, continuous improvement in all of these areas is encouraged. Annually the Executive Committee of the Center evaluates each program and project and the related strategies. The National Evaluation Committee gives particular attention to the significance of the problems being attacked and the adequacy of the strategies being employed. The University Policy Review Board examines all elements of the Center activities but gives particular attention to proper staffing. Field visitors appointed by USOE officials, and USOE officials also evaluate all components of the Center. University officials as individuals, consultants invited by the Center staff, and school people provide evaluative information. The information from all sources is carefully analyzed by all the Center staff, but particularly by the Executive Committee, Management Council, and Director. Annually, the information is formally put into program plans for the ensuing year. Changes are also made as essential at any time during the year. When any element does not function properly or produce at a reasonably high level, it is modified. The Center now demonstrates higher productivity and more sophisticated strategies than a year ago. Its improvement mechanisms are designed to accelerate these trends.

FOOTNOTE

1. Miss Elizabeth Schwenn, Project Assistant, assisted in editing this chapter. The Executive Committee and Management Council of the Center contributed to the program description of the Center and the various faculty and their research and development staff are responsible for the results as reported herein.

An Output-Oriented Model of Research and Development and Its Relationship to Educational Improvement

Hendrik D. Gideonse[1]
United States Office of Education
Washington, D. C.

FOR SEVERAL months I have followed the pages of Science attending to the continuing discussions of research and science policy, generally concerning myself with the degree to which the policy questions which arise in educational research and development can, must, or ought to be considered as a subset of those which arise for science as a whole. Thus it has been with no small degree of interest that I have read the articles prepared by Philip Handler (1967), George Daniels (1967), Harvey Brooks (1967), Lee DuBridge (1967), and Michael Reagan (1967).

Coincidental to this continuing dialogue I have been for the past 2 years engaged in a general effort to improve the quality, effectiveness, and impact of the total research effort in education. No small part of that time has been spent grappling in practice with many of the same kinds of issues as receive explication in the articles cited above. For some months I have been thinking about research and development in education trying to develop a model which would express the different functions within the total research effort, the various sources of initiative for these different kinds of activity, and the relationships among both the functions and the sources of initiative. My thinking has, of course, been in no small measure stimulated by recent debate about the ways in which schools, instruction, and education are likely to be improved most quickly and with the most substantial cumulative and lasting impact.

Reviewing the several statements cited above in terms of their general usefulness for spotlighting and clarifying analogous problems in the field of educational research revealed the particular value of Reagan's analysis as a summary of points of view already advanced and as an identification of the generally individualistic starting point of many of the theoreticians of science and research policy. The model presented in this article constitutes, I think, an instance of exactly that for which Reagan called (1967), namely, a discussion of research and development as seen and understood from the point of view of someone who sees himself as a sponsor or user of research and development to achieve certain broad purposes in a particular field of human activity. It was not developed directly in response to Reagan's suggestion; in fact, it was only after a colleague recalled his article with specific reference to the similarity of view on the research and development distinction that I realized how closely what I had formulated came to responding to Reagan's suggestion for more dialogue on the part of users of research to more adequately balance that which had been generated by the performers.

The Guba-Clark Model

Typically, the context for the debate about research and development for education is a discussion of the change process in education. Various models of change have been proposed. The one which seems to have the highest degree of currency in education at the moment places research, development, and dissemination in a linear arrangement beginning on the left with inquiry (research) and proceeding to the right through development, diffusion, and adoption (see Figure 1). In proposing the model Egon Guba

Figure 1. A classification schema of processes related to and necessary for change in education*

		Development		Diffusion		Adoption		
	Research	Invention	Design	Dissemination	Demonstration	Trial	Installation	Institutionalization
Objective	To advance knowledge	To formulate a new solution to an operating problem or to a class of operating problems, i. e., to innovate	To order and to to systematize the components of the invented solution; to construct an innovation package for institutional use, i. e., to engineer	To create widespread awareness of the invention among practitioners, i. e., to inform	To afford an opportunity to examine and assess operating qualities of the invention, i. e., to build conviction	To build familiarity with the invention and provide a basis for assessing the quality, value, fit, and utility of the invention in a particular institution, i. e., to test	To fit the characteristics of the invention to the characteristics of the adopting institution, i. e., to operationalize	To assimilate the invention as an integral and accepted component of the system, i. e., to establish
Criteria	Validity (internal and external)	Face Validity (appropriateness) --- Estimated Viability --- Impact (relative contribution)	Institutional Feasibility --- Generalizability --- Performance	Intelligibility --- Fidelity --- Pervasiveness --- Impact (extent to which it affects key targets)	Credibility --- Convenience --- Evidential Assessment	Adaptability --- Feasibility --- Action	Effectiveness --- Efficiency	Continuity --- Valuation --- Support
Relation to Change	Provides basis for invention	Produces the invention	Engineers and packages the invention	Informs about the invention	Builds conviction about the invention	Tries out the invention in the context of a particular situation	Operationalizes the invention for use in a specific institution	Establishes the invention as a part of an ongoing program; converts it to a ''non-innovation''

*From Guba and Clark, 1965

and David Clark (1965) called three caveats to attention. First, they noted that the model was constructed on logical grounds and that it was largely unsupported by empirical research. Second, they pointed out that it was not necessary for change to begin at the research or inquiry stage. Third, they noted that the model itself was a uni-dimensional analysis of change roles which are influenced by a multi-dimensional range of variables not entirely accommodated by the model's structure.

As a model of the change process, this particular schema is simple and logical. However, those of us who have worked intensely on problems of research policy in education see some shortcomings. The Guba-Clark model does not emphasize sufficiently within its structure that initiative for action of quite different kinds can take place at any point in their continuum. Because of its linear nature, and despite the second caveat mentioned above, the model unwittingly implies that innovations begin with the findings generated by fundamental research.

The Output Model

The purpose of developing an alternative model is to create a heuristic which a) illustrates the essential differences between research and development activities and b) shows how the two are – or can be – related to one another and to the operating educational system. Such a model ought to illustrate the different sources of initiative and motivation for beginning various activities. It should be able to show or imply the interplay among all the functions in the effort to improve instruction.

The model developed is based on the conviction that research, development, and school operations are different kinds of activities with quite different objectives or outputs. It indicates that initiatives for each kind of activity are the results of decisions based on quite different kinds of data and equally different kinds of internal and external needs. The model implies that while there may be a strong logical flow from the production of knowledge through the development of processes to their installation in operational settings, there may be just as strong a flow backwards as operational problems define development programs, which, in turn, reveal the need for certain basic information and theory.

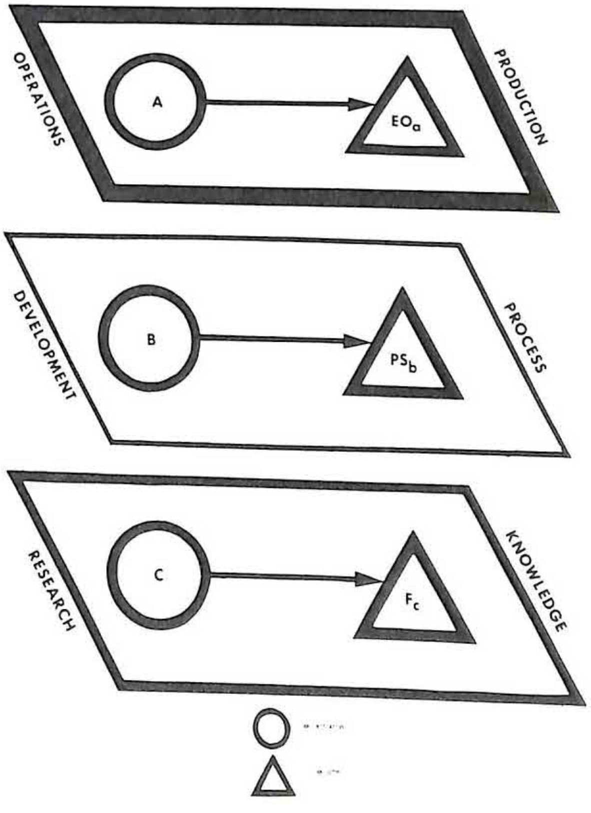
Figure 2 expresses the model. Three planes are shown, each symbolizing the different orientation of activities conducted under research, development, and school operations. (The model is – as are all models – an abstraction from reality. In the real world these activities are not always neatly separated either in time or location. The point of conceptually – and therefore graphically – separating them here is to illustrate the essentially different orientations of the three types of activity and the consequences of those differences.) For each activity represented in Figure 2, the model depicts an initiative leading to an output characteristic of that activity.

The lower plane symbolizes the knowledge orientation of research. The object of research is to generate new knowledge. One of the significant features of research is that when an activity is begun the specific outcome is not known. For research, C represents an initiative undertaken which culminates in a research finding represented by Fc.

The middle plane symbolizes what I call the process orientation of development. The object of development is to produce materials, techniques, processes, hardware, and organizational formats for instruction which accomplish certain pre-specified objectives construed to be part of the broader goals of education. One of the signficant features of development is that when an activity is begun, the objective is known or established at the outset. The objectives for a development project are cast in the form of performance specifications (PS), and all activities are geared to producing the necessary products and processes which will meet those specifications. In Figure 2, B indicates an initiative undertaken for development culminating in the creation of a process which meets performance specifications PS_b.

The top plane symbolizes the activities characteristic of school operations. The operating educational system is production oriented. Thus, the object of school operations is to act upon human beings in order to train and develop in them various skills, attitudes, beliefs, and knowledge calculated to serve both society and themselves. Certainly one of the significant features of initiatives in school operations

Figure 2. An output model of educational research and development



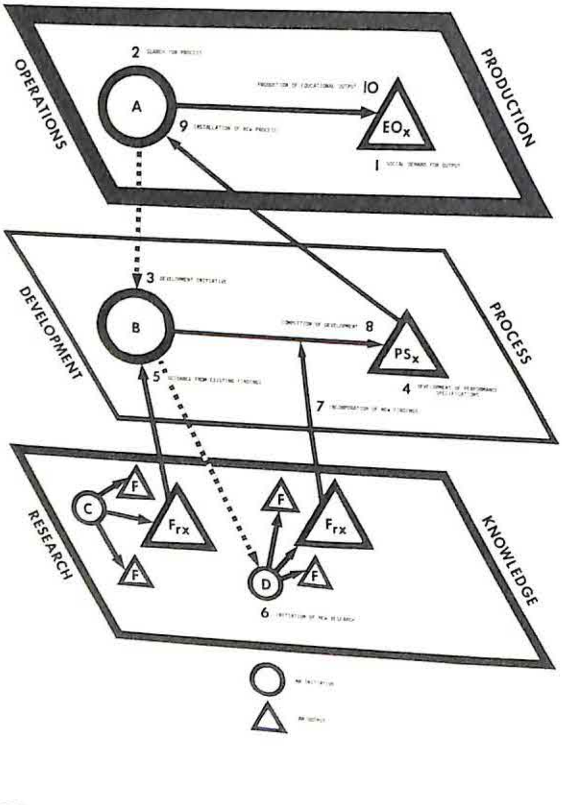
is the weight of the responsibility on the school administrator for choosing the right kinds of processes to achieve the outputs that society and individuals specify. In Figure 2, A represents an initiative to install a process leading to the production of education output EO_a.

A "Walk-Through" of the Model

To illustrate the relationships among the three types of activities, consider the following example. A responsible school official, faced with evidence that certain outputs desired by the society are not being achieved for a significant portion of the children in his charge, searches other school operations and ongoing or completed development projects for processes designed to meet his need. Should he find nothing to suit his particular problem (e. g., the low reading achievement of culturally disadvantaged children), he may then exercise his prerogative to call for the initiation of a project to design and develop a process whose performance specifications are such that upon installation of the process in his school, it will yield the desired educational output (e. g., increased level of reading achievement in the target population).

Once the initiative for the development project has been undertaken and the performance specifications established, the development project then conducts a search for relevant research findings which may offer clues to guide the development project. (Whether or not this step is taken after the project is begun or immediately before it is not really important. What is crucial is that at some point near the very beginning of the effort such a search is made.) Impressed with a particular finding (e. g., the tremendous impact of parental attitude on student achievement as revealed by the survey conducted as part of the Plowden study), the project may decide to develop a process which deliberately tries to engender a large measure of parental involvement in home instructional experiences carefully geared to complementary experiences in the school setting. Having made that decision the developers may then discover that they require further information about the specific nature of optimum parent-child interactions to stimulate maximum learner achievement. They may therefore call for a specific initiative of a research activity to generate further data to guide the development of materials. When useful findings are identified they can be incorporated in the development effort which then proceeds to successful conclusion. When, using iterative techniques of design, development, trial, and redesign on the basis of feedback, materials encompassing both school experiences and parent-child interactions in the home are successfully developed and validated, they may then be transferred to the operating setting where the administrator may install them as part of his instructional program.

Figure 3. A "Walk-Through" of the model


OPERATIONS
PRODUCTION
2 SEARCH FOR PROCESS
A
PRODUCTION OF EDUCATIONAL OUTPUT 10
9 INSTALLATION OF NEW PROCESS
EOx
1 SOCIAL DEMAND FOR OUTPUT
DEVELOPMENT
PROCESS
3 DEVELOPMENT INITIATIVE
B
COMPLETION OF DEVELOPMENT 8
PSx
5 GUIDANCE FROM EXISTING FINDINGS
4 DEVELOPMENT OF PERFORMANCE SPECIFICATIONS
7 INCORPORATION OF NEW FINDINGS
RESEARCH
KNOWLEDGE
C
F
Frx
F
F
Frx
D
F
6 INITIATION OF NEW RESEARCH
AN INITIATIVE
AN OUTPUT

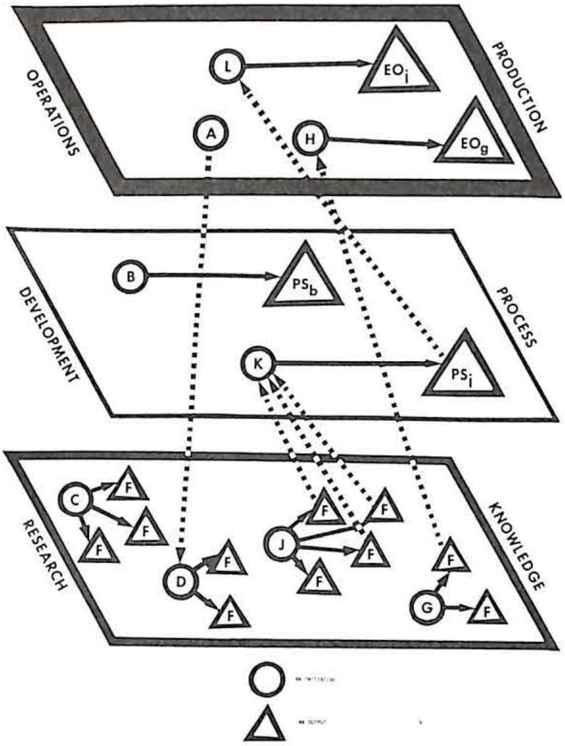
This example is illustrated in terms of the model by Figure 3. EO_x at No. 1 symbolizes the social demand for a certain kind of educational output (in the example just given, higher reading achievement for culturally disadvantaged children). This demand creates pressure on the school administrator to respond with some sort of initiative. That initiative is represented by A at No. 2. It symbolizes his search for an effective process to install. Since he did not find it, his response was to call for a development initiative (B at No. 3). The next step was to develop the performance specifications (PS_x at No. 4) such that they correspond to the educational output desired by society. Once the specifications for the development project are established, the next step is to survey related research seeking guidance for the development effort. The search is conducted and the finding (F_{rx}) of relevance to the performance specifications and the desired educational output (e. g., the significance of parental attitude) is incorporated into the development project (No. 5). The call for additional research assistance is symbolized by D at No. 6 and the incorporation of relevant findings (again F_{rx}) from that initiative into the development effort is symbolized by the solid arrow to the development line at No. 7. Number 8 represents the completion of the development project, No. 9 the incorporation of the process into school operations, and No. 10 the production of the desired output, higher levels of reading achievement, as a consequence.

A Sampling of Other Possible Interactions

The depiction of a sampling of other possible interactions among research, development, and operations can be found in Figure 4.

Example: A school official feels the need to assess the degree to which instructional programs are

Figure 4. A sampling of other possible interactions



serving a particular target population. He calls therefore for an initiative in research. This is represented by the D/D/F-F interaction.

Example: An organization engaged in development independently concludes that it would be useful to develop a certain process or product for instruction. This is represented by the B/PS_b interaction.

Example: Research is initiated for its own sake and pursued solely for the knowledge which it produces. No findings have yet been incorporated either in development or operations. This is symbolized by the C/F-F-F relation in research.

Example: Research initiated for its own sake yields the findings that certain organizational structures for large city school systems are always problematical or that a certain vitamin supplement administered between the ages of 5 and 7 can prevent a form of mental retardation whose appearance cannot be detected until somewhat later. Neither one of these findings needs to pass through development. Each can be implemented directly in school operations (if so desired). This kind of relationship of research to practice is illustrated by the G/F-F/H/EO_g interaction.

Example: Finally, consider an instance in the form of the linear flow or Guba-Clark model from research to development to implementation. Research on early childhood and cognitive growth, for example, uncovers a number of findings some of which may suggest the development of processes and environments which can actively enhance such growth. Development efforts are consequently supported and carried through to completion. The availability of the resulting products and processes is made known to educational administrators who are thereby persuaded to incorporate the newly developed early childhood instructional programs, at either private or public expense, into the Nation's educational system. This kind of sequence is represented by the J/F-F-F/K/PS_j/L/EO_j interaction.[2]

THE MODEL AS AN HEURISTIC

All of these representations in Figures 2, 3, and 4 are fairly obvious and straightforward. The use of the model as an heuristic, however, profits from further explanation. One of these uses pertains directly to the problem of "change process" as applied to education. The model is structured to illustrate that the incorporation of research findings into development is just as important and oftentimes as difficult a proposition as incorporating newly developed processes into operational settings. The notion that there are obligations on both research and development to transfer their "products" to other activities means that each must pay careful attention to the way in which its outputs are presented and, perhaps the very way in which the outputs are produced. In other words, the requirement that eventually there be transfer or incorporation into another type of activity places constraints upon the professional behavior in each activity which cannot be ignored without compromising later impact.

This requirement is particularly true for development projects, but it is as true for research activities. A few concrete examples illustrate the point. One such example is the researcher who publishes his findings in a sloppy or difficult format and thereby hinders the likelihood of their being incorporated ultimately into practice. The researcher who inadvertently conceals or compromises his methodology or design encounters similar problems.

Similarly, the final requirement for a development project is that it be usable in operational settings. The ultimate desire to incorporate the developed process in school operations means that one of the performance specifications for development must always be the provision of procedures (teachers' manuals, training procedures, etc.) for accomplishing the installation of the innovations. If the development is undertaken without reference to that fact (if, in short, the requirement for transfer is not built into the performance specifications), the developer may well have rendered his product unusable. Hence, for example, the desire to involve teachers and other practitioners in the development process stems from the need to have their expertise and experience continually represented. They constitute, in effect, the embodiment of many of the operational constraints and possibilities in schools within which the finished process will operate. There are other good reasons for involving teachers as well, not the least of them being the respectability lent to the project in the eyes of the practitioners at large by virtue of the meaningful presence of teachers in the effort. This last consideration is of no small importance in securing acceptance of the innovation in the profession at large. The nature of its importance, however, should not be mistaken; the involvement must be meaningful and not merely window-dressing, for the respectability is lent by their presence only if their contributions are fully utilized and their real knowledge of what is or may be possible is thoroughly weighed in the development effort.

One final point might be made about the model's portrayal of the possibility for transfers back and forth among research, development, and operations. That is the obvious emphasis which needs to be given to the problem of information flow and the need for carefully considering techniques for installation of better knowledge and better processes into their intended settings. Only part of this is the direct responsibility of the researchers and the developers. Those with obligations for considering the entire R and D system for improvement need also to direct their attention to the diffusion process.

A second feature of the model as an heuristic is the way it helps to clear up part of the problem of distinguishing between basic and applied research in education.[3] The model clearly implies that basic research (studies generated independently in research for the sake of the findings alone) and applied research (research conducted to serve a particular need identified by people engaged in development or operations) differ from one another primarily in terms of the intent of the initiator.[4] Thus the knowledge-orientation of the basic researcher is central to his activity. Applied research is also supported for the knowledge which results from it, but the initiator of the research knows to what instrumental use he is going to put his findings. By depicting both applied and basic research as similar kinds of activities, the model implies that *in and of themselves* they look very much alike. The procedures, the design, the sophistication must all be on par if either is to be valuable. What distinguishes the two from one another, if anything does, are the purposes for which they are initiated.[5]

A third feature of the model as an heuristic is its suggestion that decisions to initiate activities of each of the three types are made according to quite different criteria and perhaps by quite different people. The fundamental scientific character of research suggests that independent initiatives exercised there depend heavily on advice from the science community. Development projects, however, can also be independently initiated, but decisions to begin these kinds of activities are subject to advice from both research and operations. With limited resources, deciding which needs to satisfy through development (for example, those independently generated by developers, compared to those stemming directly from school operations, compared to those growing out of research activities) becomes a particularly difficult problem. Finally, the kinds of lonely decisions required of school administrators at the operational level are made by people in the context of still different circumstances and institutions. By emphasizing the essentially different nature of the activities being undertaken, the model reminds the policy maker of the need to collect different kinds of data and statements of need when planning future activities.

Finally, the frank attempt to represent each of the activities in terms of particular kinds of outputs may well be the most significant aspect of the model. It forces the viewer of the model to consider what the outputs of each activity are and to think about how the outputs of each activity are of use to one another. The outputs of research, for example, are knowledge. Some of the knowledge produced through research will find its way into development and into school operation. Are there ways of improving the output of research, making it more powerful, increase the likelihood of its being of use to instruction and education?

What about the outputs of development? They constitute, on the one hand, the validation of research and, on the other, the means by which the educational system can improve the manner in which it carries out its functions. How can development be improved, how can research be organized to be of greater use to development, and how can the educational system itself orient its organization to the recurring need for the installation of more powerful validated techniques?

Finally, what happens to educational operations when they begin to view their responsibilities in terms of output? The contrast can perhaps be most sharply drawn by considering the implications of the notion of grading *schools* on the basis of their *outputs* rather than *students* on the basis of their *performance*. The existing practice of grading *students* assumes at bottom that the student is responsible for his learning and that his failure or success is a tribute or a consequence of factors intrinsic to him. The idea of grading a *school* on the basis of its outputs assumes quite to the contrary that all students can learn and that the responsibility of the schools is to make that happen. (We do not, for example, judge the success or failure of medicine or law by the patient's or client's end state; we judge it by the degree to which the doctors or lawyers skillfully utilized the most sophisticated practices in attempting to serve the client. We certainly do not "grade" the patient or client; quite to the contrary, it is the professional services themselves which are assessed. An output orientation for school operations would cause the same reorientation of the direction of assessment in education.) If the schools themselves are judged in terms of the degree to which they are accomplishing their "production goals," increasingly they may come to orient their activities to assessing their own effectiveness, identifying the techniques and processes which need improvement and, as a consequence, calling with increasing sophistication for the kind of development activity and research support which will provide the basis for continuing improvement.

FOOTNOTES

1. The author is Director of the Program Planning and Evaluation Staff of the Bureau of Research, U. S. Office of Education. This paper is written in Gideonse' private capacity. No official

support or endorsement by the U. S. Office of Education is intended or should be inferred.

2. This example is, of course, clearly an ideal conceptualization. It is instructive to keep the model in mind as one looks back over the past four years at the tremendous developing interest in the establishment of early childhood educational programs. While it is certainly difficult if not impossible to establish an individual cause for such a complex phenomenon, it is nonetheless significant, I think, that Benjamin Bloom published his volume Stability and Change in Human Characteristics (1964) just at the time the Office of Economic Opportunity was beginning its planning toward the development of programs to fight the war against poverty. The significant thing about Bloom's volume, however, was that while his conclusions firmly underscored the importance of the early years in the development of cognitive skills, he also pointedly observed that there had been little actual experimentation designed to create environments to enhance such skill development. I do not think that it does violence to Bloom's argument to interpret it as a call for rigorous development efforts designed to produce environments and instructional programs which have the effect of enhancing human capabilities. The problem, however, is that which exists in all social domains. There are large numbers of children now whom we cannot afford to ignore, and the consequence has been the attempt to create operating early childhood programs based on those convincing research findings without first having gone through a developmental stage. The result has been a somewhat marginal impact on the target population despite the clear implications of the research summarized by Bloom.

3. I think it is also useful at this point to recall Reagan's view that the social sciences present something of a different picture in regard to the distinction between basic and applied research. As he points out (Reagan, 1967, p. 1385), no matter how abstruse and abstract its practitioners attempt to become, social science research is inherently related to potential applications.

4. The initiator and performer are, of course, not necessarily one and the same person or institution. The actual initiator of the research project may be a school man, a developer, a researcher, or a research administrator. The performer of that research effort may or may not have the same ultimate purpose as the initiator in mind as he undertakes the activity. Thus, for example, it would be perfectly possible for a research administrator to stimulate a series of research activities relating to reading which he views as applied research necessary for a development effort to build improved reading curricula, while the performer of that research sees it as a basic research effort in perception or the psychomotor skills associated with reading.

5. This view, I believe, fits fairly well with one part of Harvey Brooks' analysis of the distinctions between basic and applied research (Brooks, 1967, p. 1706) when he noted that "as definite categories, (the terms) basic and applied tend to be meaningless, but as positions on a scale within a given environment they probably do have some significance." The principal shortcoming with Brook's analysis, however, is that it proceeds almost entirely on the presumption that the distinction can be resolved by approaching it in some way from the researcher's point of view. My experience in the administration of research, as the model presented in this paper clearly indicates, leads me to believe that the researcher's view is only one of several which must be taken into account in attempting any analysis of the distinction between basic and applied research. The presence of rather different criteria and alternative vantage points convinces me that Reagan's argument for abandoning the distinction within research is much more persuasive as is the suggestion that the critical categorization is that between research and development.

REFERENCES

Bloom, B. S. Stability and change in human characteristics. New York: Wiley, 1964.

Brooks, H. Applied science and technological progress. Science, 1967, 156, 1706-1712.

Daniels, G. H. The pure-science ideal and democratic culture. Science, 1967, 156, 1699-1705.

DuBridge, L. A. University basic research. Science, 1967, 157, 648-650.

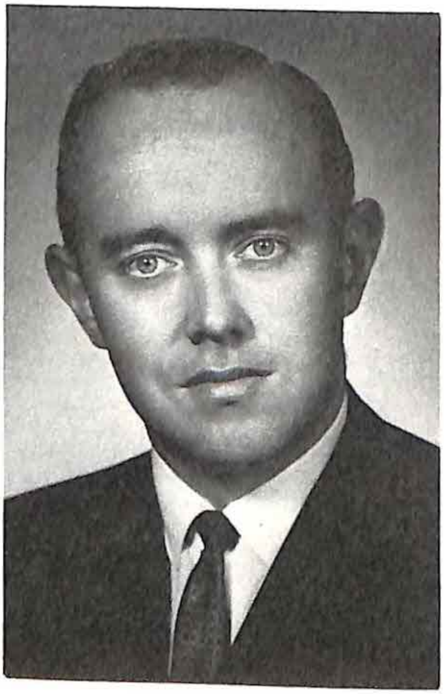
Guba, E. G., and Clark, D. L. An examination of potential change roles in education. Paper read at Symposium on Innovation in Planning School Curricula held by National Education Association Committee for Study of Instruction, Airielhouse, Virginia, October, 1965.

Handler, P. Academic science and the federal government. Science, 1967, 157, 1140-1146.

Reagan, M. D. Basic and applied research: A meaningful distinction? Science, 1967, 155, 1383-1386.

HERBERT J. KLAUSMEIER is presently Director of the Wisconsin Research and Development Center for Cognitive Learning and Professor of Educational Psychology at The University of Wisconsin at Madison. Since 1949 when he was awarded the doctorate at Stanford University, Professor Klausmeier has received recognition as a teacher, scholar, researcher, and research administrator. He is or has been an officer or national committee member of the American Psychological Association, American Educational Research Association, Association for Supervision and Curriculum Development, and the National Society of College Teachers of Education. Dr. Klausmeier is the author of several books, including the widely used *Learning and Human Abilities: Educational Psychology*. Through his numerous articles and research studies, Professor Klausmeier is widely known for his scholarship on concept learning and cognitive processes. Under his leadership the Wisconsin R & D Center has become the nation's foremost research and development organization performing basic research on processes and conditions of learning and developing related programs of instruction.

GEORGE T. O'HEARN, recently appointed Associate Professor and Acting Director of the Division of Education at The University of Wisconsin–Green Bay, has been Assistant Professor of Science Education on the UW–Madison campus. In addition to his teaching duties, Professor O'Hearn has been Associate Director of the Science Literacy Center and a Principal Investigator at the Wisconsin Research and Development Center for Cognitive Learning. A Wisconsin PhD, Dr. O'Hearn has conducted research on concept learning in science and on scientific literacy. In addition he has authored numerous articles in science education.



PERTINENT RECENT TITLES FROM DERS

THE MEASUREMENT AND PREDICTION OF TEACHER EFFECTIVENESS

By Prof. A. S. Barr et al.

144 pages. $5.

Can good teachers be distinguished from poor teachers? This milestone monograph presents an overview of 30 years of important research and offers some general observations and new hypotheses.

ABILITY GROUPING IN THE PUBLIC SCHOOLS

By Prof. Walter R. Borg

104 pages. $3.

This significant field study analyzes the generally superior learning achievements of elementary, junior high, and high school pupils in an ability-grouping system, and discusses accompanying personality problems.

YEAR-ROUND EDUCATION

By Prof. Clarence A. Schoenfeld and Neil Schmitz

144 pages. $3.

The problems and possibilities in the all-year calendar, from kindergarten to college. A unique assessment of a national issue that focuses on facts rather than on fancy.

SCHOLARS GUIDE TO JOURNALS OF EDUCATION

By Profs. L. Joseph Lins and Robert A. Rees

160 pages. $3.95.

An annotated index of 135 leading journals of education and educational psychology so authors and teachers can find out what's printed and how to communicate their findings.

THE HUMANIST TRADITION IN MODERN INDIAN EDUCATIONAL THOUGHT

By K. G. Saiyidain

256 pages. $5.95.

Here we have a very able Indian scholar writing about Indian leaders as an Indian sees them functioning in India and on the international scene. Gives Western readers an insight into Indian educational trends.

IDEAS AND IMAGES FOR LIFE WITH YOUNG PEOPLE

By Prof. Lindley J. Stiles

96 pages. $3.50.

Popular poetry about and for young people—and for all who share with youth the exciting venture of growing up.

RESEARCH AND DEVELOPMENT
TOWARD THE IMPROVEMENT
OF EDUCATION

Edited by Herbert J. Klausmeier and George T. O'Hearn. 1968

176 pages $7.25 hardcover $5.75 softcover

**DEMBAR
EDUCATIONAL
RESEARCH
SERVICES, INC.**

POST OFFICE BOX 1605 . MADISON, WISCONSIN 53701

PLEASE SEND ME THE INDICATED NUMBER OF COPIES OF:

☐ Research and Development Toward the Improvement of Education at $7.25 hardcover or $5.75 softcover.

☐ The Measurement and Prediction of Teacher Effectiveness at $5.

☐ Ability Grouping in the Public Schools at $3.

☐ Year-Round Education at $3.

☐ Scholars Guide to Journals of Education at $3.95.

☐ The Humanist Tradition in Modern Indian Educational Thought at $5.95.

☐ Ideas and Images for Life with Young People at $3.50.

Name ______________________________

Address ______________________________

City ______________________________

State ______________ Zip ____________

☐ I enclose a check for postpaid books.

☐ Bill me and I'll pay the postage.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

AN EXPERIMENT IN VISUAL LEARNING AT THE ELEVENTH-GRADE LEVEL[1]

FRANCIS M. DWYER, Jr.
The Pennsylvania State University

ABSTRACT

This study was an investigation of the effectiveness of eight types of visual illustrations used to complement oral instruction. An attempt was made to evaluate the instructional value of black and white and colored visual illustrations. Two hundred sixty-four eleventh-grade students were randomly assigned by class to one of nine treatment groups. Student learning was evaluated by means of five criterion tests designed to measure different educational objectives. Comparisons among the means of the nine treatment groups yielded significant differences. It was found that the realism continuum for still visuals is not a reliable predictor of learning efficiency; however, visual illustrations were found to significantly increase student learning on three criterion tests. Color in visual illustrations did not appear to be an important variable in increasing student learning in this study.

RECENT RESEARCH and literature indicate that too much detail in an illustration tends to reduce student learning [Broadbent (3, 4), Attneave (2), Jacobson (15), Travers (24, 25), and Dwyer (9)] and that detail in visual illustrations may be reduced so much that interpretation of the illustrations by students may cause unpredictable responses [Fleming (11)]. Apparently, there is a conflict between the visual realism continuum and results of recent research. Allen (1), Hoban (14), Schramm (22), and Dwyer (9) have stated that extensive research should be conducted on the physical characteristics inherent within pictorial illustrations which lead to increased learning and to the attainment of specific educational objectives.

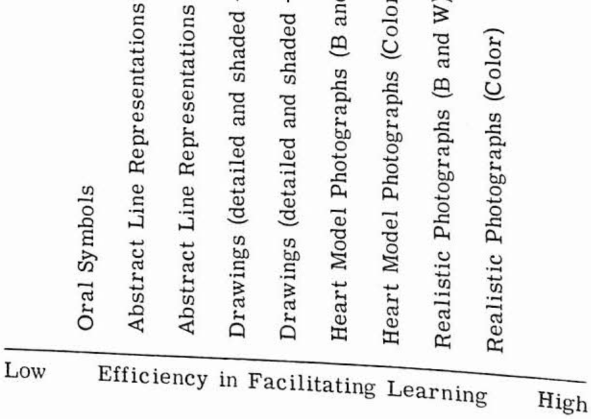
VISUAL REALISM CONTINUUM

A visual is said to be more realistic on the basis of an increase in the quantity of physical characteristics the visual has in common with the object or situation. The visual realism continuum is founded on the contention that as the amount of realistic detail in the visual illustration is increased, learning is also increased, and, thus, the visual realism continuum extends from the object or situation itself to a very simplified abstract line representation. (See Figure 1.) A visual considered to be 100 percent realistic with respect to a designated object could not be differentiated from the object because it would possess exactly the same qualities possessed by the original object. Oral instruction, therefore, would contain a low degree of realism since it would not physically represent the denota. Consequently, it would follow that visual illustrations, ranging from colored photographs to the black and white abstract line representations, may vary in their ability to promote student learning of specific educational objectives. Research based upon the visual realism continuum has been guided by the hypothesis that the higher the degree of similarity between the visual and the object or situation to which it refers, the easier it will be for the students to learn from it.

REALISM THEORIES

The present study evolved out of the inconsistencies resulting from research guided by the "established" theoretical orientations, specifically the iconicity theory identified by Morris (19), the sign similarity orientation developed by Carpenter (6), and Dale's (7) cone of experience. For purposes of

Figure 1. Realism Continuum for Still Visuals


Oral Symbols
Abstract Line Representations (B and W)
Abstract Line Representations (Color)
Drawings (detailed and shaded – B and W)
Drawings (detailed and shaded – Color)
Heart Model Photographs (B and W)
Heart Model Photographs (Color)
Realistic Photographs (B and W)
Realistic Photographs (Color)
Low
Efficiency in Facilitating Learning
High

Points on the Realism Continuum in accordance with the assumptions predicted by the "Realism Theories."

convenience the above orientations are referred to collectively as the realism theories. According to Finn (10), the basic concept around which audiovisual theories have been oriented is the concrete-to-abstract dimension of learning. The writings of several other authors appear to substantiate this basic concept presented by Finn, i. e., Gibson's (13) projective-conventional continuum, Osgood's (20) more detachable-less detachable continuum and Knowlton's (16) transparency-opacity continuum.

Although it has been established that the use of carefully prepared visual aids can facilitate student learning, there has been little attempt to determine the relative effectiveness of the various types of visual illustrations in promoting student achievement of varied educational objectives. Miller, et al. (18: 18-85) have stated that it would be a mistake to assume that one cue added to another would increase learning by a linear increment. Bruner, et al. (5) and Travers, et al. (24) have suggested that learners do not need highly embellished stimuli in order to recognize the attributes of an object or situation or to identify the object or situation. The available evidence [French (12), Miller (17), Rappaport (21), Travers (25), and Attneave (2)] seems to indicate that the effectiveness of discrimination learning promoted by the addition of relevant stimuli may be limited by the information processing capacity of the organism with learning reaching an early peak and then diminishing with the addition of relative, but superfluous, cues. Broadbent (3, 4) has explained that the reduction of learning as the cue stimulation increases is caused by the filtering process in the central nervous system which prevents much of the realistic stimuli from receiving active reception in the brain.

SPECIFIC OBJECTIVES

This study investigates the possible existence of a visual continuum representing learning effectiveness for eleventh-grade students. Specifically, this study attempts:

a. to measure the relative effectiveness of eight different types of visual illustrations used to complement oral instruction;
b. to determine at what point further increases of realistic detail in visual illustrations fail to produce significant differences in the achievement of the five specific educational objectives;
c. to compare the amount of immediate and delayed retention resulting from the use of black and white and colored visual illustrations to complement oral instruction;
d. to determine whether color in visual illustrations is an important variable in promoting student learning of the five educational objectives as measured by the criterion tests;
e. to determine whether the realism continuum proposed by the realism theories is an appropriate and reliable predictor of visual effectiveness when varied types of visual illustrations are used to complement oral instruction.

MATERIALS

It was necessary to design and photograph nine complete 2 by 2 inch slide sequences for this study. Each sequence consisted of 39 slides. Five sequences were in black and white and four were in color. The realistic photographic sequence (B and W) was the first slide sequence produced. From the realistic sequence a graphic artist reproduced the line representations and the detailed shaded drawings (both black and white and color versions). Extreme care was required to make certain that the same format and size relationship that appeared in the original heart photographs were conveyed to the drawings and eventually to the slide sequences.

The oral instruction was presented to the students by means of a tape recorder. During the oral instruction an audio signal cued the instructor to change the slide so that the presentation was synchronized. Thus, both the oral and visual instruction presented the same information simultaneously.

EXPERIMENTAL DESIGN

The subjects were 264 eleventh-grade students drawn from the Farrell School System, Farrell, Pennsylvania, and the Sullivan County High School, Laporte, Pennsylvania. The number of students involved in the delayed retention was 252; this reduction was caused by student absenteeism.

The experiment was conducted so that classes of students were assigned at random to one of the nine treatment groups. Each student in each treatment group received a pretest, participated in his particular instructional presentation, and then received four individual posttests. The students also participated in a delayed retention testing session two weeks later at which time they received the same battery of tests; they had no prior awareness of the delayed testing session.

The eight groups receiving visuals of the heart were considered to be the experimental groups and the group receiving slides on which appeared printed symbols of the terms being discussed was the control group. Information transmitted to each student via the oral channel was held constant for each presentation; the rate and number of words per minute were identical. The amount of realism possessed by the visual illustrations was varied for each of the eight experimental groups. Students viewed their respective slides for equal amounts of time.

Analysis of covariance, using scores on the 36-item physiology pretest, was used to determine the significance of difference in immediate achievement for the nine treatment groups. An analysis of covariance was also used on the data obtained from the delayed retention testing with the immediate retention total criterion test scores being used as the adjusting variable. In testing for significance of differences among the means of the several treatment groups, Dunn's c-procedure [Dunn (8); Sparks (23)] was used to compare the means of the independent samples.

CRITERION MEASURES

Drawing Test

The drawing test (N = 18 items) provided the student with a list of specific terms corresponding to the parts of the heart discussed in the oral presentation. The students were required to draw a representative diagram of the heart and place the numbers of the listed parts in their respective positions. In this test emphasis was placed on the positioning of the verbal symbols with respect to their concrete referents, not the quality of the diagram they were to draw. (Kuder-Richardson Formula 21 Reliability Coefficient: Immediate Retention .88; Delayed Retention .84.)

Identification Test

This multiple choice test (N = 20 items) required students to identify the numbered parts on a detailed drawing of a heart. Each part of the heart, which had been discussed in the oral presentation, was numbered on the drawing. The students were provided with an answer sheet with corresponding numbers on which they could select the name of the part of the heart which corresponded to the number. (Kuder-Richardson Formula 21 Reliability Coefficient: Immediate Retention .77; Delayed Retention .74.)

Terminology Test

This test (N = 20 items) consisted of a series of fill-in questions; it was possible for one and only one specific answer to be correct. This test attempted to measure the learner's knowledge of referents for specific symbols or knowledge of the referent which would be most appropriately associated with the use of a specific symbol. (Kuder-Richardson Formula 21 Reliability Coefficient: Immediate Retention .69; Delayed Retention .74.)

Comprehension Test

The comprehension test consisted of 20 multiple-choice items. Direct reference was made to the position of specific parts of the heart while functioning. The student was then asked to identify the position that other specified parts of the heart would be occupying at that particular moment. This test required that the students have a thorough understanding of the heart, its parts, its internal functioning, and the simultaneous processes which occur during the systolic and diastolic phases. The comprehension test was designed to measure a type of understanding that occurs when the individual understands what is being communicated and when he can use the information being received to explain some other phenomena. (Kuder-Richardson Formula 21 Reliability Coefficient: Immediate Retention .63; Delayed Retention .60.)

Total Criterion Test

Scores received on the four individual criterion tests were combined in a composite 78-item criterion test designed to measure total understanding of the concepts presented. (Kuder-Richardson Formula 21 Reliability Coefficient: Immediate Retention .90; Delayed Retention .91.)

Physiology Pretest

A physiology pretest consisting of 36 multiple-choice items was administered to all participants in an attempt to determine their prior factual knowledge of the functional aspects of human physiology. Scores on this test were used as the adjusting variable in the analysis of covariance technique to evaluate the relative effectiveness of the various treatment groups receiving oral instruction complemented by visual illustrations. (Kuder-Richardson Reliability Coefficient: Immediate Retention .61; Delayed Retention .90.)

TREATMENT GROUPS

Oral Presentation: Group I

Students in Group I received oral instruction throughout the entire presentation from a tape recorder. Printed word symbols depicting the proper spelling of the parts of the heart were projected on the screen as they were mentioned in the oral presentation. There were no visual illustrations used in the oral presentation.

Abstract Line Presentation (B and W): Group II

Students in Group II received the same oral instruction as students in Group I. In addition, they received abstract line illustrations (black lines on a white field) depicting the form and the relative locations of the parts of the heart as they were mentioned in the oral presentation. The line drawings used in this presentation were similar to instructional drawings used in many current science textbooks.

Abstract Line Presentation (Color): Group III

Students in Group III received the same oral instruction as students in Group I. In addition, they received abstract line visual illustrations (blue lines on a pink field) similar to those received by students in Group II; however, these illustrations were in color.

Detailed, Shaded Drawing Presentation (B and W): Group IV

Students in Group IV received the same oral instruction as students in Group I. In addition, they viewed detailed, shaded drawings representing the parts of the heart as they were mentioned in the oral instruction. These drawings were more complex than the abstract line drawings and more realistically represented the heart and its related parts rather than merely identifying and locating them as in the former line representations.

Detailed, Shaded Drawing Presentation (Color): Group V

Students in Group V received the same oral instruction as students in Group I. In addition, they received visual illustrations similar to the illustrations received by the students in Group IV; however, these visuals were in realistic color.

Heart Model Presentation (B and W): Group VI

Students in Group VI received the same oral instruction as students in Group I. In addition, they viewed photographs of a heart model as the heart and its related parts were being discussed orally.

Heart Model Presentation (Color): Group VII

Students in Group VII received the same oral instruction as students in Group I. In addition, they viewed colored photographs of a heart model as the heart and its related parts were being discussed orally.

Realistic Photographic Presentation (B and W): Group VIII

Students in Group VIII received the same oral instruction as students in Group I. In addition, they viewed realistic black and white photographs of the heart as the heart and its related parts were being described orally.

Realistic Photographic Presentation (Color): Group IX

Students in Group IX received the same oral instruction as students in Group I. In addition, they viewed realistic colored photographs of the heart similar to the black and white sequence received by students in Group VIII.

SUMMARY OF ANALYSES: IMMEDIATE RETENTION

On the total criterion test and on three of the four individual criterion tests, significant differences were found to exist among the means of the nine groups. The F ratios for the adjusted analyses were all significant beyond the 0.01 level with the exception of the terminology test where no significant differences were found to exist.

A. On the total criterion test the various slide sequences facilitated total student learning with differing degrees of effectiveness.

  1. The abstract line presentations (B and W and Color), the detailed, shaded drawing presentation (Color), and the realistic photographic presentation (B and W) were more effective than the oral presentation alone.

  2. The oral presentation alone was as effective as the detailed, shaded drawing presentations (B and W), the heart model presentation (B and W and Color), and the realistic photographic presentation (Color).

  3. When the black and white and color versions of each visual presentation were compared in terms of their instructional effectiveness, no significant differences were found.

B. In relation to the promotion of student achievement on the drawing test:

  1. The abstract line presentations (B and W and Color) and the detailed, shaded drawing presentation (Color) were more effective than the oral presentation alone.

  2. The oral presentation alone was as effective as the detailed, shaded drawing presentation (B and W), the heart model presentations (B and W and Color) and the realistic photographic presentations (B and W and Color).

  3. When the black and white and color versions of each visual presentation were compared in terms of their instructional effectiveness, no significant differences were found.

C. In relation to the promotion of student achievement on the identification test:

  1. The abstract line presentation (B and W), the detailed, shaded drawing presentation (Color), and the realistic photographic presentation (B and W) were more effective than the oral presentation alone.

  2. The oral presentation alone was as effective as the abstract line presentation (Color), the detailed, shaded drawing presentation (B and W), the heart model presentations (B and W and Color), and the realistic photographic presentation (Color).

  3. The detailed, shaded drawing presentation (Color) was more effective than the detailed, shaded drawing presentation (B and W), and the heart model presentations (B and W and Color).

D. In relation to the promotion of student achievement on the terminology test, The F-ratio did not exceed the critical value for a 0.05 level test. Further analysis among the various group means was not justified.

E. In relation to the promotion of student achievement on the comprehension test:

  1. The oral presentation alone was as effective as each of the visual presentations.

  2. The abstract line presentations (B and W and Color), the detailed, shaded drawing presentations (B and W and Color), and the realistic photographic presentation (B and W) were more effective than the realistic photographic presentation (Color).

INTERPRETATION OF ANALYSES: IMMEDIATE RETENTION

In terms of instructional effectiveness, economy, and simplicity of production:

1. The abstract line presentation (B and W) should be used to promote total student understanding of the concepts presented in the instruction.

2. The abstract line presentation (B and W) should be used to promote student learning of specific locations of the various patterns, structures, and positions of the parts of the heart.

3. The abstract line presentation (B and W) should be used to promote student transfer of learning, i. e., the ability to identify the

numbered parts on a diagram of the heart from information presented via the oral-visual presentation.

4. The oral presentation alone should be used to promote student learning of the referents of terms.

5. The oral presentation alone should be used to promote the development of a new view, or reorganization, of the material by the student.

SUMMARY OF ANALYSES: DELAYED RETENTION

In relation to the promotion of student achievement on the total criterion test, the drawing test, the identification test and the comprehension test, the F-ratio did not exceed the critical value for a 0.05 level test. Further analysis among the various group means was not justified. Significant differences were found to exist among the means on the terminology test (0.05 level).

A. In relation to the promotion of student achievement on the terminology test:

   1. The oral presentation alone was as effective as each of the visual presentations.

   2. When the black and white color versions of each visual presentation were compared in terms of their instructional effectiveness, no significant differences were found.

INTERPRETATION OF ANALYSES: DELAYED RETENTION

In terms of instructional effectiveness, economy, and simplicity of production, the oral presentation alone should be used to promote student achievement of the five types of educational objectives measured in this study.

CONCLUSION

Emerging from the study are a number of generalizations for the use of visual illustrations for promoting the achievement of specific learning objectives in the biological sciences by eleventh-grade students. Further research is needed, however, to determine whether the same types of visual illustrations are equally effective in promoting student learnings of subject matter in different content areas and for students in different grade levels.

1. The visual realism continuum is not a reliable predictor of learning efficiency for eleventh-grade students. The findings of this study seem to substantiate the hypothesis that reality may be edited for instructional purposes. This would seem to raise a number of important questions regarding the instructional value of many types of visual aids being produced commercially.

2. Visual illustrations used to complement oral instruction significantly increase student achievement of the educational objectives measured in this study by the total criterion test, the drawing test, and the identification test.

3. In terms of instructional effectiveness, economy, and simplicity of production the abstract line representations (B and W) should be used to complement eleventh-grade instruction in this study by the total criterion test, the drawing test and the identification test.

4. Color in visual illustrations presented to eleventh-grade students does not appear to be an important variable in increasing student learning of educational objectives measured in this study.

5. Although it was found that the presentations using visuals were just as effective as the oral presentation in facilitating student achievement on the delayed retention testing, extreme caution is necessary in the interpretation of this finding. Since the retention tests were administered two weeks after the immediate retention session, further research is needed where the time lapse between the immediate and delayed testing sessions is much greater (four, eight, twelve, etc.) before any definite recommendations can be made.

FOOTNOTE

1. This article is based on research sponsored by the U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research. Complete details of the procedures described can be found in the Final Report Project No. 6-8840, Grant No. OEG-1-7-068840-0290 by F. M. Dwyer, Jr., which includes this and other studies.

REFERENCES

1. Allen, W. H. "Audio-Visual Communication." In C. W. Harris (Ed.) Encyclopedia of Educational Research, Third Edition (New York: Macmillan, 1960).
2. Attneave, F. Application of Information Theory to Psychology (New York: Henry Holt, 1959).
3. Broadbent, D. E. Perception and Communication (New York: Pergamon Press, 1958).
4. Broadbent, D. E. "Information Processing in the Nervous System." Science, 1965, 3695, 457-462.
5. Bruner, J. S., Goodnow, J. J., and Austin, G. A. A Study of Thinking (New York: John Wiley and Sons, 1956).
6. Carpenter, C. R. "A Theoretical Orientation for Instructional Film Research." Audio-Visual Communication Review, 1953, 1, 38-52.
7. Dale, E. Audio-Visual Methods in Teaching (New York: Dryden Press, 1946).
8. Dunn, O. J. "Multiple Comparisons Among Means." American Statistical Association Journal, 1961, 56, 52-64.

9. Dwyer, F. M. "Adapting Visual Illustrations for Effective Learning." Harvard Educational Review, 1967, 37, 250-263.

10. Finn, J. D. "Professionalizing the Audio-Visual Field." Audio-Visual Communication Review, Vol. I, Winter 1953: 6-17.

11. Flemming, M. L. An Experimental Exploration of Three Communicator Strategies for Increasing the Predictability of Pictorial Cue – Verbal Response Relationships. Unpublished doctoral dissertation, Indiana University, 1960.

12. French, R. S. "Pattern Recognition in the Presence of Visual Noise." Journal of Experimental Psychology, 1954, 47, 27-36.

13. Gibson, J. J. "A Theory of Pictorial Perception." Audio-Visual Communication Review, 1954, 2, 3-23.

14. Hoban, C. F. "The Usable Residue of Educational Film Research." In Institute for Communication Research (Eds.), New Teaching Aids for the American Classroom (Stanford: Stanford University, 1960).

15. Jacobson, H. "The Information Capacity of the Human Eye." Science, 1951, 113, 292-293.

16. Knowlton, J. Q. A Conceptual Scheme for the Audio-Visual Field. Bulletin of the School of Education, Indiana University, Vol. 40, May 1964.

17. Miller, G. A. "The Magical Number Seven, Plus or Minus Two; Some Limitations on Our Capacity for Processing Information." Psychological Review, 1956, 63, 81-97.

18. Miller, N. E., et al., (Eds.) "Graphic Communication and the Crisis in Education." Audio-Visual Communication Review, 1957, 5, 1-120.

19. Morris, C. W. Signs, Language, and Behavior (New York: Prentice-Hall, 1946).

20. Osgood, C. E. Method and Theory in Experimental Psychology (New York: Oxford University Press, 1953).

21. Rappaport, M. "The Role of Redundancy in the Discrimination of Visual Forms." Journal of Experimental Psychology, 1957, 53, 3-10.

22. Schramm, W. F. New Teaching Aids for the American Classroom (Stanford, California: Institute for Communication Research, 1960).

23. Sparks, J. N. "Expository Notes on the Problem of Making Multiple Comparisons in a Completely Randomized Design." Journal of Experimental Education, 1963, 4, 344-349.

24. Travers, R. M. W., McCormick, M. C., Van Mondfrans, A. P., and Williams, F. E. Research and Theory Related to Audio-Visual Information Transmission (Utah: Bureau of Educational Research, University of Utah, 1964).

25. Travers, R. M. W. "The Transmission of Information to Human Perceivers." Audio-Visual Communication Review, 1964, 12, 373-385.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

DELIBERATIVE ATTITUDES, IQ, AND SOCIAL CLASS

T. BENTLEY EDWARDS
University of California, Berkeley

ABSTRACT

When large numbers of children are polled concerning their attitudes toward school and academic study, those from poor homes are, on the average, less likely to hold favorable attitudes than those from better homes. This fact is hardly surprising. But the relationship between favorable attitudes and socio-economic status is far less marked than is the relationship between IQ and socio-economic status. Hence, when children are selected for high-ability classes and eventually for college on the basis of IQ or related measures, a child who likes ideas is far more likely to be rejected by the school system if he comes from the lower socio-economic classes. The foregoing conclusion is based on a survey of thousands of children in elementary and high schools in this country and in England.

ATTITUDES FAVORABLE TO DELIBERATION

Elsewhere, the development of scales to measure attitudes favorable to the acquisition of understanding needed for deliberation prior to the making of choices has been described (2). Attitudes favorable to the study of things as opposed to people have been separated and labeled Theoretic as opposed to Prudent. The fate of children with Theoretic interests has been discussed (3). Although possession of either Theoretic or Prudent interests was shown to be advantageous for school achievement, children with Prudent attitudes received disproportionately more of school rewards, that is high grades, and those with Prudent as opposed to Theoretic interests were found more likely to be selected for advanced educational opportunities.

In the present article, data will be presented to show that although scores on the so-called Intelligence Tests correlate highly with socio-economic status, a fact which has been established by data from many investigations (5, 4, 1, 6), scores on attitude tests do not. There is a correlation, but it is much less pronounced. Hence, when children and youth are selected for advanced education on the basis of intelligence test scores, or on other test scores or marks that correlate highly with intelligence test scores, many children with strong deliberative interest scores are rejected. These days, when concern for equality of educational opportunity is so urgently voiced, fresh tests are needed with which to distinguish those members of the less favored social classes, many of whom are Negroes, who might respond to a more enlightening school environment than the one they now encounter.

SOURCES OF DATA

Attitude tests were administered to 3,750 students in 12 high schools around the San Francisco Bay. Intelligence test scores, and data on occupation of father were collected for the same students.

The 12 schools were chosen so as to provide as wide a range of intelligence and occupation of father as possible rather than to provide a representative sample. To this end, census data were analyzed and schools selected that seemed likely to produce generous samples of rural and urban residents; upper, middle, and lower social classes; inhabitants of comfortable suburbs, of industrial slums, and of pockets or urban wealth; and of Negro and of white members of the various socio-economic classes. The assumption taken here is that when such socio-economic variations are sampled, the full range of attitude and of intelligence test scores is likely to be included.

To facilitate data collection, schools within driving distance of Berkeley were used, and only those schools eager to cooperate in the study were included. Descriptions were written for each of the 60 odd high schools within a 50-mile radius of the University. These schools were divided into four major groups according as they were found in areas classed by the Bureau of the Census as: metropolitan, urbanized areas, cities over 5,000 and places under 5,000. In the first of these, three large metropolitan schools were selected more or less arbitrarily to include an all Negro school (School No. 11, Table 1), School No. 3 with representation from several social classes, and the school (No. 13) used in a pilot study. A unique private school (No. 2) in the metropolitan area was also included. This school is heavily endowed and is practically tuition-free. According to the terms of the endowment, all those attending the school must take courses in the manual and industrial arts shops along with their academic courses.

Attempts were made to achieve a representative

TABLE 1

DISTRIBUTION OF FATHERS' OCCUPATIONS

Number	Community Setting	Manual Labor	White Collar	Small Business and Mer.	Farm Owner	Profes- sional	Execu- tive
1	Town--rural marketing center	42%	20%	19%	6%	7%	1%
2	Metropolitan--boy's private school	6	24	18	0	30	21
3	Metropolitan--heterogeneous	38	33	16	1	8	4
4	City--light industry	52	23	6	4	12	3
5	Town--rural	43	20	12	19	5	1
6	Town--rural	33	32	17	8	7	3
7	City--industrial	47	29	17	0	6	1
8	Urban--working class residential	55	32	11	0	1	1
9	Suburban	23	36	17	0	16	8
10	Urban--industrial	67	18	12	0	3	0
11	Metropolitan--Negro residential	74	14	6	0	6	0
12	Urban--upper class residential	2	19	19	0	25	35
13	Metropolitan--"white collar" residential	29	33	25	1	12	0
All Schools Combined		40%	27%	16%	3%	9%	5%

TABLE 2

DISTRIBUTION OF FATHERS' EDUCATION

Fathers' education terminated during or at conclusion of:	School Number													All Schools Combined
	1	2	3	4	5	6	7	8	9	10	11	12	13	
Elementary school	25%	3%	19%	28%	46%	17%	37%	28%	11%	23%	32%	0%	11%	22%
High School	51	18	49	48	35	53	49	57	42	60	40	20	45	45
College	19	44	21	13	13	17	9	9	34	8	16	50	31	21
Post-graduate	5	34	9	8	4	11	4	3	12	0	4	26	10	9
Not available	0	1	2	3	2	2	1	3	1	9	8	4	3	3

sample in the Negro high school and the heterogeneous school. In the pilot study school, almost the entire population of Grade 11 was tested, as was the entire population from the private school. The total number of students selected from each area was chosen so as to match as nearly as possible the proportions of the population of the State of California living in such areas.

Tables 2, 3, and 4 show the distributions of fathers' education, IQ scores of the students, and their racial origins. In schools No. 2 and No. 12, more than 75 percent of the fathers are college graduates. Over half of these fathers are in the professional and executive classes. For school No. 2 six percent, and for school No. 12 two percent of the fathers work with their hands. In each of these two schools one father in five operates a small business. None are farmers. In school No. 2, 31 percent, and in school No. 12, 39 percent of the children show IQ's in excess of 120. In school No. 2 only one child in a hundred shows an IQ of less than 100. In school No. 12, 13 percent show an IQ of less than 100. School No. 2 is 97 percent white; school No. 12 is 100 percent white.

By way of contrast, 46 percent and 37 percent of the fathers in schools No. 5 and No. 7 did not attend high school. Between 40 percent and 50 percent of these fathers work with their hands. Nearly a third of the children in these schools show an IQ of 100-110. In school No. 5, 96 percent, and in school No. 7, 86 percent of the children are white. The first

TABLE 3

DISTRIBUTION OF IQ TEST SCORES

IQ Scores	School Number													All Schools Combined
	1	2	3	4	5	6	7	8	9	10	11	12	13	
150 plus	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	0%
140-149	0	1	1	0	0	1	1	0	0	1	0	2	2	1
130-139	1	6	3	2	1	3	1	1	2	0	0	10	6	2
120-129	6	24	12	3	3	4	7	4	7	9	1	26	11	8
110-119	21	43	20	15	17	15	23	12	16	9	8	27	24	18
100-109	27	25	22	9	38	21	31	16	22	28	22	16	21	20
90-99	23	1	15	10	21	15	21	14	16	23	24	11	12	15
80-89	12	0	14	3	5	12	6	7	5	23	25	2	7	9
70-79	5	0	7	2	2	1	1	2	2	1	17	0	2	4
60-69	0	0	1	0	0	1	1	1	0	0	1	0	0	1
Below 60	0	0	1	0	0	0	0	1	0	0	0	0	0	0
Not Available	5	0	4	56	13	27	8	42	30	6	2	5	14	22
Test	Mix	Otis	T-M	CMM	CMM-p	CMM	Otis	T-M	Otis	CMM	T-M	T-M	T-M	

TABLE 4

RACIAL DISTRIBUTION

Race	1	2	3	4	5	6	7	8	9	10	11	12	13	All Schools Combined
Negro	0%	1%	29%	1%	0%	1%	10%	1%	0%	22%	91%	0%	7%	13%
Oriental	0	2	10	0	4	2	3	1	0	4	2	0	6	3
White	100	97	60	97	96	96	86	97	99	71	5	99	86	83
Not Available	0	0	1	2	0	1	1	1	1	3	2	1	1	1
Number of Cases	388	160	380	227	135	141	222	538	288	123	311	96	601	3750

of these two schools is in a small town in the country; the second is in a small heavily industrialized town.

Schools No. 8 and No. 10 are in heavily industrialized communities within the metropolitan district. Fathers of children in No. 8 are nearly all in the working classes; 55 percent work with their hands; 32 percent are white collar workers, and 11 percent own small businesses. Fathers of children in No. 10 show 67 percent, 18 percent, and 12 percent in these three categories. Although 57 percent and 60 percent of these fathers graduated from high school, only 12 percent and 8 percent are reported as college graduates. Only 60 percent of the IQ's from school No. 8 were available, but these show a fairly normal distribution. Those from school No. 10, however, show 23 percent from 90-99 and 23 percent from 80-89. In school No. 8, 97 percent of the children are white. In school No. 10, 71 percent are white.

In one other school, No. 11, 91 percent of the children are Negroes. Only one child per hundred, in this school, shows an IQ over 120, and less than 10 percent show IQ's in excess of 110. Three-quarters of the children show IQ's ranging from 80-109, and 17 percent show IQ's of 70-79. Fathers of 40 percent of these children are reported as having graduated from high school, and 20 percent are reported as college graduates. Many of the children are sons and daughters of workers attracted from the south during World War II. None of these fathers is an executive, but 6 percent are reported as being in the professional class. Three-quarters of the fathers are manual workers.

INTELLIGENCE TEST SCORES

All schools do not give the same tests of intelligence. In the 12 schools selected, the Terman-McNemar was given most frequently, but some of the schools gave the Otis, and others gave the California Test of Mental Maturity.

Correction factors are available, or could be worked out to make scores on these various tests closely comparable, but because only gross comparisons were to be made in the present study, correction factors have not been applied. The necessary information is available and the increased accuracy may be sought at a later date.

So far as the present study is concerned, the correlations indicated would probably be increased slightly if the data on IQ were more accurate. No change in the conclusions drawn could be anticipated.

FATHERS' OCCUPATION

The students themselves were asked to report their fathers' occupation. Attached to the attitude inventory was a personal data sheet which included the following item:

> 22. My father's (or step-father's) job or occupation is ____________
> Just what does your father do in his job? ____________

The explanation called for in the second part of the question helped to decide whether the child really knew what his father did at his work. Further corroboration was supplied when the child was asked about his father's education. Unless the various responses were in agreement, the datum on occupation was not used. An elaborate key was prepared with which to classify the father's occupational status. This key includes the orientation of the occupation according to the four poles of the attitude inventory, but the orientation of the occupation does not enter the argument of the present article which refers only to the socio-economic status of the father's occupation. The complete key is provided in Table 5.

ATTITUDES

The attitude inventory has been published (2), but for convenience several sample items from the Prudent-Immediate, Theoretic-Immediate, and Prudent-Theoretic scales are reproduced in Table 6.

Both Prudent and Theoretic orientations are assumed to indicate a liking for mental deliberation prior to choice-making. Prudent also means an orientation toward the social environment, Theoretic toward the non-social. Hence a person scoring high on Prudent (1 - 2) would be demonstrating a liking for those ideas about people encountered in the study of history, political science, sociology, anthropology, psychology, and literature. A person scoring high on Theoretic (3 - 4 on the P - T scale) would be demonstrating a liking for those ideas about things encountered in the study of natural science, mathematics, and engineering. Immediate interests provide the needed contrast and demonstrate a liking for more immediate and proximate clues in the solution of life's daily problems. Conversation, personal experience, and intuitive feeling are more likely to govern the choices made than are the use of principles and modes of thought learned in studying formal school subjects or disciplines.

ASSOCIATIONS INDICATED

Tables 9 - 12 show the association of Prudent or Theoretic as opposed to Immediate orientations with IQ. Tables 7 and 8 show Prudent as opposed to Theoretic orientations and IQ. These are offered by way of contrast. Scores are tabled separately for boys and girls, and separation into three groups according to fathers' occupation is also provided.

Tables 9 - 12 show that although IQ is linked closely with occupation of father, a deliberative attitude, whether Prudent or Theoretic, is less closely linked with either IQ, or occupation of father. This means that when children are selected for advanced educational opportunities by means of IQ scores, the children of working class parents are at a disadvantage whether they have deliberative attitudes or not. This is also true when children are selected for high ability groupings by means of IQ scores or by other test scores that correlate highly with IQ. The foregoing is based on Prudent-Immediate and Theoretic-Immediate scores. Prudent-Aesthetic and Theoretic-Aesthetic scores lead to the same conclusions.

Attitude scores can vary only from one through four. It is important to remember that by the scoring technique, approximately one quarter or 25 percent of the population studied was arbitrarily assigned each of the four ranks. The percentage of Prudent or Theoretic in each of the various classifications should normally be 45-55, and departure from this norm must be offset by a similar departure in another grouping. For example, 44 percent of the boys are Prudent as opposed to Theoretic. But 76 percent of the girls are Prudent. Among 200 students, half boys and half girls, we should therefore find 120 (76 plus 44) who are Prudent as opposed to Theoretic. The reason for the large number of Prudent as opposed to Theoretic is that the cut-off points were decided on the basis of responses from a pilot group which was more Theoretic than the more general sample. The resulting imbalance in no way interferes with the argument of the present article.

By contrast, 47 percent of the boys and 58 percent of the girls are Prudent as opposed to Immediate. Hence among 200 students, we should find 105 (47 and 58) who are Prudent.

The number of 100 in any of the cells, unless the population of the cell were small, would mean heavy depletion in some other cell. For example, 91 percent of the high IQ daughters of professionals are Prudent on the P-I scale. This is offset by 30 percent for the low IQ boys. The total percentage of Prudent for the girls is 58. This is balanced by a total of only 47 for the boys. For the Bay Area study as a whole, 1,315 out of 2,518, or 52 percent show Prudent scores, when all those scoring one

TABLE 5

ORIENTATION OF OCCUPATION

Occupational Status	1 - Prudent	2 - Theoretic	3 - Aesthetic	4 - Immediate	y - Not Classifiable
1 - Manual: semi- and unskilled			machine operator, truck driver, janitor, gardner, domestic, "laborer," seamstress, hod carrier, mechanic's helper, etc.		Army private, PFC, Cpl.; Navy seaman
2 - Manual: skilled			journeyman: printer, welder, sheet metal worker, electrician, painter, carpenter, radio-TV repairman, baker, forester		Army tech and staff sgts.; Navy petty officers
3 - Lower white collar	nurse's assistant, practical nurse, camp counselor	Radiation Laboratory: electronics technician, radar technician	postman, fireman, bus driver, professional athlete, fireman, dance instructor, model, beautician, copy draftsman, photo technician, laboratory assistant	sales clerk, bookkeeper, stenographer secretary, union official, foreman, policeman, guard, stewardess, waiter, bar tender, driver-salesman (e.g., bread, milk), inspector	Army master or 1st sgt, warrant, Navy chief petty officer, warrent-bosun, Coast Guard mate
4 - Upper white collar	public school teacher, nurse, librarian, social welfare worker, probation officer, city playground director, adult education	pilot, laboratory technician, high school science teacher, chemical technician, accelerator technician (Rad Lab)	music teacher, dress designer, commercial artist, cartoonist, dance band musician, design draftsman, fish and game warden. PE teacher, dental tech, dietician, materials designer	salesman, store manager, chief of police, buyer, purchasing agent, office manager, claims adjuster, FBI, police captain, ass't sales manager	Army lieutenants, captain, Navy chief warrant, ensign, lieutenants
5 - Self Employed: Artisan			mechanic, barber, shoemaker, nurseryman, taxidermist, plumber, carpenter-contractor, grinding-shop owner, jeweler (watch repair), gardner, dog-trainer, piano technician		

TABLE 5 (Continued)

Occupational Status	1 – Prudent	2 – Theoretic	3 – Aesthetic	4 – Immediate	y – Not Classifiable
6 – Self Employed: Merchant				retail store owner (e.g., grocery, variety), export-import business, broker, druggist, contractor, mortician	
7 – Self-Employed: Farmer			farmer, rancher, horse breeder, or trainer		
8 – Professional: Salaried	professor of social science or humanities, judge, minister, missionary, editor, journalist, criminologist, high school principal	professor of natural science or mathematics, engineer, scientist, medical research, research forester, aeronautical engineer	professor of music, art, or architecture	chief accountant and statistician for a group of insurance companies	Army major, colonels, Navy commanders, captain
9 – Professional: Self-Employed	doctor, psychiatrist, lawyer, author		architect, dentist, veterinarian, automobile designer	certified public accountant	
0 – Executive	senator, representative, governor, president (of USA)			sales manager, factory representative, corporation executives, sales engineer, production engineer, management consultant, university comptroller, registrar	general, admiral
x – Miscellaneous			farm laborer		x – x: housewife
y – Not Classifiable	"working to make peace in the world," etc.	"Radiation Lab," "work with minerals," "electronics"	artist, musician, actress, TV or movie star, ballet dancer, violinist, tap dancer, champion ice skater, organist	"businessman," "office work"	no response, "don't know," not classifiable (e.g., name of a company), student, unemployed, "army," "navy," "WAC," "WAVE," deceased, retired (former occupation not given)

TABLE 6

SAMPLE ITEMS

Prudent-Immediate Scale	Theoretic-Immediate Scale	Prudent-Immediate Scale
Science has definitely not been able to show that colored races are inferior to white races.	I should rather be elected to the Student Council than be selected as an honor student in science.	I am more interested in finding out how TV has affected people's taste than in finding out how TV works.
A business man should make his decisions strictly according to the interests of his business. He should not worry about what happens nationally to wages and prices.	I would rather be known as the writer of a social column published in many papers than as the Director of an astronomical observatory.	I am more interested in following newspaper reports on the recent discoveries regarding "negative matter" than on the developments in racial integration in the schools.

TABLE 7

BAY AREA--BOYS P - T (PERCENTAGES PRUDENT BY IQ AND FATHER'S OCCUPATION)

Occupation of Father	IQ					Total
	120+	110-119	100-109	90-99	-89	
Professional and executive	49	39	45	27	37	43
N	84	88	58	11	8	249
Percentage of class in each IQ bracket	34	36	23	4	3	100
White collar	43	50	36	40	41	43
N	82	137	129	82	41	471
Percentage of class in each IQ bracket	17	29	28	18	8	100
Manual worker	45	40	40	40	62	45
N	58	120	174	141	138	631
Percentage of class in each IQ bracket	9	19	28	22	22	100
Totals	45	44	40	39	56	44
N	224	345	361	234	187	1351
Percentage in each IQ bracket	17	25	27	17	14	100

and those scoring two on the P-I scale are added together. This means that 100 - 52, 48 percent of the same group of students are classified as Immediate.

We are now ready for a more detailed study of Tables 7 - 12.

PRUDENT-THEORETIC

Attention to percentage of class in each IQ bracket shows the association of IQ with social class. More than a third, 34 percent, of the sons of the professional and executive class have an IQ in excess of 120. By contrast, only 9 percent of the sons of manual workers are in this group showing an IQ in excess of 120. Nearly half, 44 percent, of the sons of manual workers show an IQ below 100. Only 7 percent of the sons of the professional and executive class are below 100. The percentages for the girls are similar, 34, 9; 44, and 15.

Attention to the percentage of Prudent in each IQ

TABLE 8

BAY AREA--GIRLS P - T (PERCENTAGES PRUDENT BY IQ AND FATHER'S OCCUPATION)

	IQ					
Occupation of Father	120+	110-119	100-109	90-99	-89	Total
Professional and executive	82	76	65	94	33	78
N	45	42	26	17	3	133
Percentage of class in each IQ bracket	34	32	20	13	2	100
White collar	84	78	75	68	74	76
N	74	111	133	75	50	443
Percentage of class in each IQ bracket	16	25	30	17	11	100
Manual worker	80	87	75	68	78	76
N	59	121	179	155	129	643
Percentage of class in each IQ bracket	9	19	28	24	20	100
Totals	82	80	74	70	76	76
N	178	274	338	247	182	1219
Percentage in each IQ bracket	14	23	28	20	15	100

TABLE 9

BAY AREA--BOYS P - I (PERCENTAGES PRUDENT BY IQ AND FATHER'S OCCUPATION)

	IQ					
Occupation of Father	120+	110-119	100-109	90-99	-89	Total
Professional and executive	61	59	53	36	16	56
N	84	88	58	11	6	247
Percentage of class in each IQ bracket	34	36	23	5	2	100
White collar	52	50	43	41	30	45
N	82	137	129	82	33	463
Percentage of class in each IQ bracket	18	29	28	18	7	100
Manual worker	67	49	45	38	31	44
N	58	120	174	141	96	589
Percentage of class in each IQ bracket	10	21	29	24	16	100
Totals	59	52	46	39	30	47
N	224	345	361	234	135	1299
Percentage in each IQ bracket	17	27	28	18	10	100

TABLE 10

BAY AREA--GIRLS P - I (PERCENTAGES PRUDENT BY IQ AND FATHER'S OCCUPATION)

	IQ					
Occupation of Father	120+	110-119	100-109	90-99	-89	Total
Professional and executive	91	69	58	59	67	73
N	45	42	26	17	3	133
Percentage of class in each IQ bracket	34	32	20	13	2	100
White collar	78	80	60	55	48	60
N	74	111	133	75	50	443
Percentage of class in each IQ bracket	16	25	30	17	11	100
Manual worker	64	64	56	48	40	53
N	59	121	179	155	129	643
Percentage of class in each IQ bracket	9	19	28	24	20	100
Totals	77	65	55	51	42	58
N	178	274	338	247	182	1219
Percentage in each IQ bracket	14	23	28	20	15	100

TABLE 11

BAY AREA--BOYS T - I (PERCENTAGES THEORETIC BY IQ AND FATHER'S OCCUPATION)

	IQ					
Occupation of Father	120+	110-119	100-109	90-99	-89	Totals
Professional or executive	64	82	71	64	50	71
N	84	88	58	11	8	249
Percentage of class in each IQ bracket	36	36	23	5	2	100
White collar	63	63	63	56	59	62
N	82	137	129	82	41	471
Percentage of class in each IQ bracket	18	29	28	18	7	100
Manual worker	71	63	59	59	61	62
N	58	120	174	141	136	629
Percentage of class in each IQ bracket	10	21	29	24	16	100
Totals	66	68	63	59	60	64
N	224	345	361	234	185	1349
Percentage in each IQ bracket	17	27	28	18	10	100

TABLE 12

BAY AREA--GIRLS T - I (PERCENTAGES THEORETIC BY IQ AND FATHER'S OCCUPATION)

	IQ					
Occupation of Father	120+	110-119	100-109	90-99	-89	Total
Professional and executive	49	55	62	24	33	50
N	45	42	26	17	3	133
Percentage of class in each bracket	34	32	20	13	2	100
White collar	34	41	44	36	46	40
N	74	111	133	75	50	443
Percentage of class in each bracket	16	25	30	17	11	100
Manual workers	49	45	45	46	45	46
N	59	121	179	155	129	643
Percentage of class in each bracket	9	19	28	24	20	100
Totals	43	44	45	40	45	44
N	178	274	338	247	182	1219
Percentage in each IQ bracket	14	23	28	20	15	100

bracket for each social class shows that boys generally are more likely to be Theoretic than Prudent. The percentage varies around a mean of 44. Among the upper class boys, half the high IQ's are Prudent, but of the low IQ's only about a third are Prudent. Among the lower class boys half the high IQ's are Prudent, but among the low IQ's two-thirds are Prudent. This reversal among the low IQ sons of manual workers is surprising and may indicate that the boys as a group are reading something into the items that other groups do not. Or it may mean that this group of boys, faced with failure on all sides, is laying claim to attitudes which do not correlate with their phenomenal lives. For this group of boys the position taken as they respond to the items may be less realistic than for other groups; a substantial number may be answering from the world of fantasy to which their education and experience is forcing them to retreat. Boys with IQ's below 90, especially if they are the children of manual workers, are not likely to be welcomed into shop or laboratory any more than into those classes where verbal skills are paramount. In turn, they may know enough about science and technology to repudiate them, and to have been deceived about the concerns of literature and history by their experiences in low ability classes.

Three of every four girls are Prudent and the Proportion varies little with socio-economic class or with IQ. For the boys, the Prudent-Theoretic proportions do vary with social-class but the change is far less marked than the variation in IQ with social class. For the girls, there is almost the same marked variation in IQ with social class but almost no variation in Prudent-Theoretic orientation.

The point to be made is that orientation of attitude is much less firmly linked with social class than is IQ. So far as the Prudent-Theoretic contrast is concerned, no severe loss of deliberative orientation can be attributed to selection on the basis of IQ or on the basis of tests that correlate with IQ, since Prudent and Theoretic orientations are both deliberative. But when attention is paid to scales contrasting deliberative and non-deliberative attitudes, similar absence of correlation with social class produces an unexpected and disturbing result when children are selected on the basis of IQ.

PRUDENT-IMMEDIATE

For boys, the association between Prudent and high IQ is about the same as it is for girls but for neither sex is the association as marked as is the link between high IQ and high socio-economic status. For example, 61 percent of the sons of members of the professional and executive class with an IQ above 120 have Prudent as opposed to Immediate tastes, whereas for sons of professionals with an IQ between 90 and 99 (chosen because the number below 90 was too small) the proportion is 36 percent. For girls the figures are 91 percent and 59 percent. But the proportion of boys with an IQ in excess of 120 is three and a half, 34 to 10, when the percentages among the professionals and manual workers is compared. For girls, the proportions are almost the same.

This incongruity is produced by the fact that the link between social class and a Prudent orientation is light. Slightly more than half, 56 percent, of the sons of professionals are Prudent as opposed to Immediate. But 44 percent of the sons of manual workers are also Prudent. For girls, the link between social class and a Prudent orientation is greater than for boys but still not nearly as marked as is the link between social class and high IQ. Nearly three-quarters, or 73 percent of the daughters of professionals are Prudent; just over half, or 53 percent of the daughters of manual workers are Prudent.

THEORETIC-IMMEDIATE

The Theoretic-Immediate comparisons with IQ and social class show somewhat similar relationships except that the differences between boys and girls are reversed and the association of a Theoretic preference with IQ is less marked. In fact, a Theoretic as opposed to Immediate preference is as likely to be found among the low IQ girls as among the high IQ girls and almost as likely among the low as among the high IQ boys. More than 60 percent of boys of all social classes show a Theoretic preference. Among girls, this preference is expressed by between 40 and 50 percent.

SELECTION BY IQ

Tables 13 and 14 show the results of the associations that have been described in a more striking fashion. For both boys and girls the two columns are nearly identical. Only one pair of columns will be described, the other three pairs of columns providing strong support for the argument. Consider data for boys in the first two columns of Table 13. When boys are selected for school rewards leading to further educational opportunities upon a basis of IQ scores, 66 percent of the sons of the professional class were so selected, but only 28 percent of the sons of manual workers. Because the sons of manual workers are as likely to indicate a preference for Prudent as opposed to Immediate activities as are the sons of professionals and executives, 72 percent of the Prudent among the sons of the professionals will be selected but only 34 percent of the Prudent among the sons of manual workers.

Table 15 uses a cut-off point of 120, a more realistic figure where the educational opportunities leading to the professions is concerned. Now, half the deliberative sons of professionals are being turned down, but four-fifths, or 80 percent, of the sons of manual workers with deliberative orientations are being refused.

Where selections have been made either on the basis of IQ or on the basis of measures that correlate highly with IQ a number of potential failures are known to have been admitted. The alleged stability of the IQ derives largely from its manufactured high relationship with past achievement. The IQ will predict accurately only for those children who continue as they have in the past. Selection of children below the arbitrary IQ cut-off point who have strong deliberative orientations, especially among the culturally deprived groups, might turn out to be a successful practice, as well as being more desirable on humane and compassionate grounds.

TABLE 13

PERCENTAGES SELECTED FROM TOTAL NUMBER OF PRUDENT (AS OPPOSED TO IMMEDIATE) CHILDREN IN EACH SOCIAL CLASS, GIVEN A CUT-OFF POINT OF IQ = 110

Father's Occupation	Boys		Girls	
	IQ over 110	Prudent	IQ over 110	Prudent
Professional and executive	66	72	70	74
White collar	41	49	47	53
Manual worker	28	34	30	37

TABLE 14

PERCENTAGES SELECTED FROM TOTAL NUMBER OF THEORETIC (AS OPPOSED TO IMMEDIATE) CHILDREN IN EACH SOCIAL CLASS, GIVEN A CUT-OFF POINT OF IQ = 110

Father's Occupation	Boys		Girls	
	IQ over 110	Theoretic	IQ over 110	Theoretic
Professional and executive	70	71	66	68
White collar	47	48	41	39
Manual worker	30	30	28	28

TABLE 15

PERCENTAGES SELECTED FROM TOTAL NUMBER OF BOYS WITH EXTREME DELIBERATIVE SCORES IN EACH SOCIAL CLASS, GIVEN A CUT-OFF POINT OF IQ = 120

Father's Occupation	Prudent P - I		Theoretic T - I	
	IQ over 120	Extreme P-score	IQ over 120	Extreme T-score
Professional and executive	34	50	34	37
White collar	18	41	18	26
Manual worker	10	19	10	10

REFERENCES

1. Benedict, Ruth, and Gene Weltfish, The Races of Mankind (Washington: Public Affairs Press, No. 85, 1946), p. 19.

2. Edwards, T. Bentley, and Alan B. Wilson, "The Development of Scales of Attitudinal Dimensions," The Journal of Experimental Education, XXVIII (September, 1959), pp. 3-36.

3. Edwards, T. Bentley, "Theoretic Interests and the Schools," Journal of Experimental Education XXXVI, No. 3 (Spring, 1968), pp. 43-51.

4. Eells, Kenneth, et al., Intelligence and Cultural Differences (Chicago: University of Chicago Press, 1951).

5. Haggard, Ernest A., "Social Status and Intelligence," Genetic Psychology Monographs, XLIX (1954), pp. 141-186.

6. Warner, W. Lloyd, Robert J. Havighurst, and Martin B. Loeb, Who Shall be Educated (New York: Harper and Brothers, 1944), pp. 239, 466.

PERTINENT RECENT TITLES FROM DERS

THE MEASUREMENT AND PREDICTION OF TEACHER EFFECTIVENESS
By Prof. A. S. Barr et al.
144 pages. $5.

Can good teachers be distinguished from poor teachers? This milestone monograph presents an overview of 30 years of important research and offers some general observations and new hypotheses.

ABILITY GROUPING IN THE PUBLIC SCHOOLS
By Prof. Walter R. Borg
104 pages. $3.

This significant field study analyzes the generally superior learning achievements of elementary, junior high, and high school pupils in an ability-grouping system, and discusses accompanying personality problems.

YEAR-ROUND EDUCATION
By Prof. Clarence A. Schoenfeld and Neil Schmitz
144 pages. $3.

The problems and possibilities in the all-year calendar, from kindergarten to college. A unique assessment of a national issue that focuses on facts rather than on fancy.

SCHOLARS GUIDE TO JOURNALS OF EDUCATION
By Profs. L. Joseph Lins and Robert A. Rees
160 pages. $3.95.

An annotated index of 135 leading journals of education and educational psychology so authors and teachers can find out what's printed and how to communicate their findings.

ders

**DEMBAR
EDUCATIONAL
RESEARCH
SERVICES, INC.**

POST OFFICE BOX 1605 . MADISON, WISCONSIN 53701

PLEASE SEND ME THE INDICATED NUMBER OF COPIES OF:

☐ The Measurement and Prediction of Teacher Effectiveness at $5.
☐ Ability Grouping in the Public Schools at $3.
☐ Year-Round Education at $3.
☐ Scholars Guide to Journals of Education at $3.95.

Name ______________________________
Address ______________________________
City ______________________________
State ______________ Zip __________
☐ I enclose a check for postpaid books.
☐ Bill me and I'll pay the postage.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

AN ELEMENTARY SCHOOL FORM OF THE DOGMATISM SCALE

RUSSELL L. FIGERT JR.
American Education Publications
Middletown, Connecticut

ABSTRACT

The Dogmatism Scale for adults was adapted to provide an instrument for use in studies on the organization of belief-disbelief systems of children in grades four-six. The paper-and-pencil scale contains 50 items of the forced-choice type. It can be administered to groups. A summated-rating scoring system is used to determine a respondent's placement along an openminded-closedminded continuum. When administered to 485 pupils in grades four through six in all five elementary schools in a central Indiana city of about 7,000 population, the instrument's reliability coefficient was found to be .67. All items discriminated between low and high scorers at P< .05.

MILTON Rokeach has suggested that studies of ways in which individuals organize beliefs and disbeliefs will lead to a better understanding of human behavior than currently exists and to implications for persons involved in the teaching process (5). He suggested that an individual possibly possesses millions of beliefs and disbeliefs, defined as hypotheses that the person consciously or unconsciously accepts as true or rejects as false. He further suggested that an individual organizes such beliefs and disbeliefs into systems having describable and measurable properties and observable behavioral consequences, and that such systems may be measured along an openminded-closedminded continuum through a sampling of that person's reactions to a series of statements.

Thus, a person having a strong belief to the effect that "the United States and Russia have just about nothing in common" may tend to accentuate differences and minimize similarities, an indication that the individual may be more closedminded than openminded (5:61-62). Rokeach postulated a number of such properties of belief-disbelief systems. He then wrote and collected statements designed to tap cognative aspects of such properties (5:73-80).

The statements were compiled into a forced-choice, summated-rating type of scale designed to measure an individual's placement along the open-closed continuum. A respondent assigned one of six possible responses, ranging from "strongly agree" to "strongly disagree," to each item. Differing responses were assigned differing numerical weights for scoring purposes.

Rokeach and other investigators used scale scores as the criterion to form groups of individuals presumed to be either high or low along the openminded-closedminded continuum. Performances of such extreme groups on tasks other than responding to items but also posited to follow from theoretical postulations were compared (5:99-410). To the extent that a group's performance of such tasks agreed with its scale-score placement along the open-closed continuum, theoretical postulations were considered valid and the instrument was considered useful as a measuring device.

In a series of investigations it was found, for example, that groups measured as openminded by the scale, in comparison to groups measured as closedminded, usually were better able to synthesize a series of beliefs needed to solve problems based on a fictitious miniature cosmology (5:171-284). Persons having such ability, Rokeach reasoned, would have belief-disbelief subsystems possessing greater degrees of intercommunication, a proposed property of belief-disbelief systems, than would persons not having such an ability.

Such a hypothetico-deductive approach to theory validation, logicians insist, requires supporting evidence from many investigations conducted in a wide variety of situations. Most of the reported studies on Rokeach's postulations have involved college students and other adults. Pannes adapted the scale and conducted a study with junior and senior high school pupils (4). A further adaptation of the scale, for use with children in grades four through six, is the subject of the remainder of this discussion (3).

SCALE ADAPTATION

The development of a form of the Dogmatism Scale for use in studies of belief-disbelief systems of children in grades four through six involved three phases of refinement. In the first two phases, items of the scale for adults were rewritten with the intention of reducing vocabulary and concept loads. The rewritten items, along with original items, were submitted to juries of reading and psychology experts for evaluations of parallelism and to establish content validity. The reading level of rewritten items, as determined through use of the Spache formula (6) and the Dale word list (2), was found to be at about grade 3.0.

The response schema for adult scale items, involving selection of one of six degrees of agreement, was changed to a system involving three degrees: agree (+), uncertain (0), or disagree (-). In scoring, a numerical weight of one was assigned to disagree responses, two to uncertain responses, and three to agree responses. Since all items were written so that agreement was a presumption of closedmindedness, the higher a respondent's score, the more closedminded he was presumed to be.

Rewritten items were given trials with pupils enrolled in grades four through six at schools in central Indiana towns. Sample populations were described as predominantly white, middle class, Protestant, native born and of northern European descent. Children were given the impression that their responses to scale items would be anonymous. On a title page, respondents were asked to supply information on their sex, age, and date of birth, but not their names. Data from school records were matched with that supplied by respondents to identify individuals. Teachers identified pupils whom they considered to be incapable of reading the items and data on such pupils were eliminated from consideration.

A pool of 84 items discriminating at $P < .05$ between high and low groups as distinguished by scale scores was obtained from the first two phases of the adaptation. In the selection of 50 items for the third phase of the instrument's development, consideration was given to selecting items representative of the various properties of belief-disbelief systems as postulated by Rokeach. Items were assigned randomly to positions in the scale.

FINDINGS

The 50-item scale (Appendix) was administered by its developer to 485 pupils in all 17 sections of grades four through six in all five schools in a city of about 7,000 population in central Indiana. Thirty-four pupils were absent. Four of the five schools were public and one was parochial. The sample population, once again, was described as being predominantly white, middle class, Protestant, native born, and of northern European descent.

Data on 27 of the 485 pupils were eliminated from analytic aspects of the study; 26 were judged by teachers to be unable to read the items and one pupil failed to mark responses to any of the 15 items on page 4 of the scale booklet. The remaining sample consisted of 230 boys and 228 girls. On a 5-point socio-economic scale, none of the pupils was judged by teachers to be in class one, the highest. Seven were assigned to class two, 160 to class three, 260 to class four, and 31 to class five, the lowest. The mean age of the population was 132 months and the range was 111 to 166 months. From information found in school records, the mean intelligence quotient of the sample population was found to be 106, with a range of 75 to 135 and a standard deviation of 11. The average achievement grade level of pupils for selected school subjects was 5.3, with a range of from 2.3 to 9.3. Pupils' scores on the instrument being developed ranged from 74 to 140, with a mean of 108 and a standard deviation of 10. The distribution of scores did not vary significantly from normal.

The median time taken by pupils to respond to the instrument was about 15 minutes and the range was from about 10 minutes to about 30 minutes. Odd-even reliability coefficients, as corrected for scale length, were found to be .73 for grade four, .60 for grade five, .70 for grade six, and .67 for the total sample. Responses to individual items by pupils in high and low 27 percents of the total distribution of scores were compared through use of a chi-square procedure and 48 items were found to discriminate at $P < .01$ and two (numbers 10 and 17) were found to discriminate at $P < .05$.

Scores attained by pupils on the revised instrument were compared by grade, socio-economic class, and school groupings by use of a critical ratio technique. The mean of scores for grade four (110.64) was found to be significantly higher ($P < .01$) than the means for both grades five and six. The mean for grade five (107.53) was higher than the mean for grade six (106.68), but not significantly so. Means by socio-economic class did not differ significantly but there was a tendency toward an inverse relationship. The mean for pupils attending the parochial school was significantly higher ($P < .05$) than for two of the four public schools and the mean for pupils attending a public school in a low socio-economic area of the city was significantly higher than the mean for pupils in a higher socio-economic section of the city.

Point biserial or product moment correlations were used to compare pupils' scores on the instrument and their sexes, ages, intelligence quotients, and achievement test scores. All coefficients were found to be in the chance range ($-.20$ to $+.20$) and none reached the $P < .05$ level of significance.

DISCUSSION AND CONCLUSIONS

Some of the data, it was concluded, could be interpreted as evidence of the instrument's validity. Assuming a relationship between grade in school and amount of knowledge possessed, the inverse order of means of scores by grade groups would seem to support Rokeach's theoretical postulation concerning a positive relationship between amount of knowledge possessed and openmindedness (5:38). In agreement with findings of other studies (1:572), there were tendencies, though not consistent or significant, for pupils in higher socio-economic classes to score lower than pupils in lower socio-economic classes. The finding concerning the higher mean for pupils in the parochial school was consistent with findings by Rokeach and others in studies of adult populations (5:117-119). None of the data, it was concluded, could be interpreted as evidence that would tend to invalidate the instrument.

For the sample population, it was concluded, the instrument was functioning relatively effectively as a measuring device and was measuring some of the same facets of openmindedness-closedmindedness among children that adult forms of the instrument measure among adults. The instrument, it was further concluded, should be given additional trials with sample populations having socio-economic, religious, and ethnic backgrounds varying from those of this study. It should also be used to select groups of high and low scorers for scale-validation studies using some of the techniques developed by Rokeach and others. Finally, results of studies should be analyzed for implications concerning the organization and presentation of instructional materials.

REFERENCES

1. Berelson, Bernard, and Steiner, Gary A. Human Behavior: An Inventory of Scientific Findings (New York: Harcourt, Brace and World, Inc. 1964).
2. Dale, Edgar. "A Comparison of Two Word Lists," Educational Research Bulletin, X (1931) 484-489.
3. Figert, Russell L. Jr. An Elementary School Form of the Dogmatism Scale, unpublished doctoral dissertation (Muncie, Indiana: Ball State University Library, 1965).
4. Pannes, Ernestine D. An Exploratory Study of the Relationship Between Attitudes Toward Self and the Degree of Dogmatism Present in a Group of Junior-Senior High School Students, unpublished doctoral dissertation (New York University, 1962).
5. Rokeach, Milton. The Open and Closed Mind (New York: Basic Books, Inc., 1960).
6. Spache, George. "A New Readability Formula for Primary-Grade Reading Materials," Elementary School Journal, VIII (1953) 410-413.

APPENDIX

Instructions and Items for the Elementary School Form of the Dogmatism Scale

WHAT DO YOU THINK ?

This is a way to find out how boys and girls think about things at different ages and in different grades. Boys and girls in grades four, five, and six are being asked to mark answers on what they think about the 50 statements on the next three pages.

The statements are what boys and girls might think or say about things. You may find that most of the statements say things in the way you would. Or you may find that only a few of the statements say things in the way you would. In any case, you will find that many boys and girls mark the statements in the same way you do. The answers you mark should be what you think about things.

The answers you give are secret. Do not put your name on your paper. Just put the grade you are in, the date of your birthday, if you are a boy or a girl, and how many years old you are.

I am in grade____. I am ____years old. I am a boy ☐ girl ☐

My birthday is on ______________ (month) ________ (day).

Here is how to mark the statements:

1. If you find that a statement says what you think is so all of the time, make a circle around the plus sign (+) to the right of it. + 0 -
2. If you find that a statement does not say what you think is so, make a circle around the minus sign (-) to the right of it. + 0 -
3. If you find that a statement says what you think is so some of the time, and does not say what you think is so some of the time, make a circle around the zero (0) to the right of it. + 0 -

Please make sure you mark one answer, and only one answer for each statement. There is no time limit, but most boys and girls are able to finish in about 30 minutes. If you need help with any of the words, raise your hand. Now go ahead with the statements.

1. Some people are always for what is true. All the rest are always against what is true. + 0 -
2. We must believe what important people say. If we do not, we will not know what is going on in the world. + 0 -
3. Sometimes I start talking before someone else is done because I must tell what I have to say. + 0 -
4. I just cannot stop when I talk with people who do not see things like I do. + 0 -
5. Most people just do not care about others. + 0 -
6. It is not good to read books that get your thoughts about things mixed up. + 0 -
7. It is wrong to give in any to those who think different people should run our country. + 0 -
8. Some ways of thinking are really the same even though people try to tell you they are different. + 0 -
9. There is no use to take time to listen to someone who will just try to change your mind. + 0 -
10. People know more about things they believe in than things they are against. + 0 -
11. In times like these, people are selfish if they think first about their own happiness. + 0 -
12. I must make myself do a lot of things that I need to do. + 0 -
13. By saying things over and over you can be sure people know what you mean. + 0 -
14. Many things in books are not worth being there. + 0 -
15. People who think about themselves first are terrible. + 0 -
16. Most people seem smart enough. Even so, groups do dumb things at times. + 0 -
17. There is so much to do and so little time to do it in. + 0 -
18. Most people just do not know what is good for them. + 0 -
19. People must take chances if they really want to do big things. + 0 -
20. It seems like many people I talk to do not really know about the good and bad things that are going on in the world. + 0 -
21. It does not matter much if you are not happy now. It is what will happen in years to come that counts. + 0 -
22. Most people seem to have odd ideas about a lot of things. + 0 -
23. When people are mad, they often bring up things that do not have much to do with what they are mad about. + 0 -
24. It is bad to make people do things. Yet, often this is the only way to help the world. + 0 -
25. I am sure people talk about me. + 0 -
26. Sometimes we must make people do things we know are good for them. + 0 -
27. People who will never say they are wrong make me mad. + 0 -
28. It is better to be a dead hero than a live "chicken." + 0 -
29. Farm people and city people think about most things in different ways. + 0 -
30. Thinking about what might happen makes people afraid. + 0 -

31. Many times I do not listen to what people are saying because I am thinking of what I will say next. + 0 -

32. People who do not believe in something important do not have much of a life. + 0 -

33. People get the most out of life only when they try hard to do what they think is best. + 0 -

34. I would like it if someone would tell me how to take care of my problems. + 0 -

35. People do not say much about it, but they are thinking about how they can become great. + 0 -

36. We have a good way of running our country. Even so, it would be better if we let only smart people do it. + 0 -

37. Most people seem to find things wrong with what others think. + 0 -

38. If people knew what I really thought, they might not like me. + 0 -

39. It is best to find out what smart people say about something before you say anything yourself. + 0 -

40. If people had the chance, they would do something great to help the world. + 0 -

41. People seem to think that most of the things they do are bad. + 0 -

42. Things are really bad when someone talks against people who think the same as he does. + 0 -

43. We are going against our own side if we listen to what the other side says. + 0 -

44. Life is terrible if people do not have help. + 0 -

45. People should not try to work together if they believe different things. + 0 -

46. It would be better if we could go back to the way things were in the good old days. + 0 -

47. It seems like people I do not even know are always looking for the bad things in me. + 0 -

48. There are many ways to think about things in this world. Even so, there is only one right way. + 0 -

49. At times I think I am no good at all. + 0 -

50. I cannot stand some people because of the way they think about things. + 0 -

---

Before you hand in your paper, please be sure you have marked one answer, and only one answer for each statement.

THANK YOU !

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

A HELPING RELATIONSHIP EXPERIENCE IN TEACHER EDUCATION[1],[2]

NEWELL T. GILL
ROY KING
RON G. WILBURN
Eastern Kentucky University

ABSTRACT

The helping relationship experience was explored as a means of insuring greater rapport with children. A sample of 82 junior and senior education students participated for a minimum of two hours/week for approximately eight weeks. Of this group, some were involved with white children; another group with Negro children; and a third group served as control. Groups were pre- and post-tested on selected personality variables felt to be necessary for the establishment of good rapport. A complex analysis of variance design was used in analyzing the data. Females and ''liberals'' scored significantly higher on the MTAI than males or ''traditionalists.'' Failure to find other evidence of a significant effect on the experimental groups was attributed to insufficient duration and/or intensity of the treatment.

IN THE National Advisory Council's report to the President on the Education of the Disadvantaged Children, November 30, 1966, following an extensive summary of 86 school districts in 43 states, the Council abstracted six main conclusions, one of which dealt directly with rapport. The factor most frequently observed that distinguished favorable classrooms from poor was the quality of the relationship between the teacher and the child (4). This would seem to underscore an urgent need for a new look at our present teacher education system. It may be that the current list of assumed characteristics felt to be necessary for an effective teacher, as currently demonstrated by the curriculum requirements in our teacher education institutions, is incomplete and/or inappropriate.

It was concluded in an earlier study by one of the investigators that the student-teacher's lack of sensitivity to children inhibited his attempt to communicate with them (3). Students subsequently were encouraged to interact with a child in a helping relationship (9) for a minimum of two hours per week. Both the student, child, and parent-teacher responses to this experience were highly favorable. All parents and teachers without exception expressed a sincere desire that the program be continued. Many of the students stated that the tutoring or helping relationship experience had been the first meaningful thing encountered while in college. Many also reported a gain in self-confidence and self-esteem. In support of this finding, others have suggested that almost any human interaction that can be classified as a helping relationship will tend to facilitate the psychological growth of all those involved (1, 2, 5). For this reason the investigators first became interested in exploring the effect of the helping relationship with children as a rather unique experience in teacher education.

THE PROBLEM AND PURPOSE OF THE STUDY

An assumption of this study is that our teacher education programs, in general, are presently not designed to adequately develop and strengthen the personality characteristics needed to insure the future establishment of good rapport between the student and his pupils. It may be agreed that openness, sensitivity, receptivity, compassion, understanding, etc. are personality traits most frequently found in individuals demonstrating the ability to establish good rapport. The problem is one of discovering means for effectively developing and strengthening such personality traits. The purpose of this study was to explore the effects of a helping relationship experience on selected personality variables felt to be significant in the establishment of good rapport.

RESEARCH HYPOTHESES

Hypothesis I	The involvement of a student with a child in a helping relationship enhances the student's general feeling and sensitivity for children.
Hypothesis II	The involvement of a student with a child in a helping relationship enhances the student's feelings about himself.
Hypothesis III	The involvement of a student with a child in a helping relationship significantly affects the ''dogmatism-liberal'' characteristics of his belief system.
Hypothesis IV	Students in a helping relationship

with Negro children tend to view themselves and the Negro as being more alike than students who have not tutored Negroes.

Hypothesis V Student performance on the Minnesota Teacher Attitude Inventory is significantly affected by the degree to which the student is 1) "dogmatic" and 2) "secure."

DESIGN OF THE STUDY

Sample

Three teachers of a child study unit at Eastern Kentucky University volunteered their students as subjects for this research. All of the students were either juniors or seniors enrolled in the teacher education program during the 1967 spring semester. The sample involved 32 white males and 50 white females.

Procedure

The three teachers who volunteered their students for this study did not completely agree on the value of the helping relationship experience. It was on this disagreement that the researchers attempted to capitalize. Two of the teachers felt such an experience should definitely be offered their students; however, only one made it a requirement. Consequently, in the case where it was optional, a few students (three males and seven females) chose an experience other than working with a child. The third teacher did not provide for the helping relationship experience.

The initiative of the students who participated in the helping relationship and the availability of the children were probably the greatest determiners as to the particular age, sex, and race of a child. Consequently, the researchers initially were not sure as to whether the subjects would distribute themselves adequately for analysis. The statistical design was thus dependent to a large degree upon the student's choice at the beginning of the semester and the availability of the children. It was the hope of the investigators that each subject be classifed as having been involved in a helping relationship with either a Negro child, a white child, or none at all. The students actually distributed themselves across the three categories in the following manner:

1) Thirty-four Ss did not participate in a helping relationship, the control group.
2) Sixteen Ss entered into a helping relationship with a Negro child, an experimental group.
3) Thirty-two Ss entered into a helping relationship with a white child, an experimental group.

All of the subjects were pre-tested during the first three weeks of the spring term prior to the initiation of the treatment and post-tested the last three weeks following the treatment. Four variables were explored in investigating the effects of the helping relationship experience, 1)sensitivity to children, 2) general security, 3) dogmatism, and 4) prejudice.

The instruments used to investigate each of the variables are listed below:

1. Minnesota Teacher's Attitude Inventory[3]
2. Maslow's Security-Index Inventory
3. Rokeach's Dogmatism Scale
4. Bogardus Social Distance Scale (Appendix A)
5. A Trait Evaluation Scale (Appendix B).

Unfortunately the MTAI was not available for pretesting. It should be recognized, therefore, that any differences found in performance on the MTAI between the experimental and control groups may have existed prior to the treatment. The other tests, however, were available from the beginning. Consequently, differences which have taken into account the relative position of each group from the start can more reasonably be attributed to the effects of the treatment.

The Treatment

The treatment consisted of a helping relationship experience (9) with a child for a minimum of two hours per week for ten weeks. The child involved in this experience was selected by the student. There were no limitations on exact age, sex or race, only that they be children. The situation (or condition) under which this relationship could function was limited only by guardian approval and the safety of the child. Most of the students in the experimental group found themselves either acting as a big brother, a baby sitter, or a tutor. Regardless of the situation, their purpose was all the same; to become a friend to the child with whom they were involved. This required a special effort on their part to become more open-minded, more accepting, more transparent, and more sensitive. It was expected that some would be more successful than others. The important aspect of the experience, however, was the fact that all of the students involved in a helping relationship were presented with an excellent opportunity to promote their own personal adequacy.

The Instruments

The Minnesota Teacher Attitude Inventory (MTAI). This scale has successfully identified teachers capable of establishing good rapport with students. Reasonably high validity and reliability coefficients have been reported. The scale has been found to be only slightly susceptable to attempts to "fake good" (Test Manual).

Maslow's Security-Index Inventory (S-I Inventory). The validity of this measure depends to a large extent upon the subject's willingness to disclose himself. Consequently, the test was not administered until it was felt that an acceptable degree of confidence had been established between the professor and his students. Of those subjects randomly selected from the sample, none felt the interpretation of their score to be invalid (Test Manual).

Rokeach's Dogmatism Scale (Form E). Five of the original statements of Form E were omitted (Nos. 11, 12, 13, 14, and 15 on Form D) from use in this study because of their similarity to the type used in Maslow's S-I Inventory. As this lowered the number of items to 35, five suitable questions were chosen from Form D (No. 4, 7, 29, 56, and 57) in order that the sample means be comparable with the norms (10).

The subjects were asked to indicate the extent to which they agreed or disagreed with each statement on a scale ranging from one to seven. The statements were all of a dogmatic nature so that strong agreement – 7 – would be interpreted as a dogmatic response and strong disagreement – 1 – as a liberal response. The theoretical mean was 160.

Social Distance Scale. The scale was made up of statements selected on an a priori basis to elicit responses indicative of the subject's degree of acceptance of the person being considered (Appendix A). Reviewers report that it has been widely and successfully used for this purpose (6).

Trait Evaluation Scale. The trait evaluation scale consisted of 24 trait dimensions (Appendix B). Each dimension was represented by two diametrically opposed traits separated by a seven-point continuum. Subjects were asked to rate someone or themselves on each of the 24 dimensions. The average of their total ratings yielded a theoretical mean position in space. Logically, it would seem that people can judge others only in relation to themselves, and furthermore, that those judged as more similar are generally perceived as more acceptable. With this in mind, subjects were asked to rate themselves and another. The difference between these scores was assumed to be representative of a theoretical distance between the subject and the other person.

Statistical Analysis

As the purpose of the study was to determine the effect on students in teacher education of a helping relationship experience, the statistical design was oriented toward exploring differences between groups in degree of performance change on tests related to selected personality variables. Differences between pre- and post-test performances were computed. As this difference score was subject to a plus or minus sign, a suitable constant was added. The results were then subjected to a complex analysis of variance to investigate possible differences in changes in performance between various groups due to the treatment (8). This procedure was appropriate for all the variables except that measured by the Minnesota Teachers Attitude Inventory. As this test was not available for pre-testing, the results obtained at the termination of the treatment could not as readily be ascribed to the effect of the treatment.

RESULTS

Coefficients of reliability were obtained on the instruments used in this study by computing a Pearson r correlation between pre- and post-test. The findings are shown in Table 1.

TABLE 1

RELIABILITY COEFFICIENTS USING TEST-RETEST METHOD

Source	N	r_{xx}
S-I Inventory	82	.88
Dogmatism Scale	82	.77
Bogardus Social Distance Scale on Negro	82	.67
Trait Evaluation Scale on Negro	82	.76
Trait Evaluation Scale on Self	82	.51

All correlations listed are significant at or beyond the .01 level.

All of the measures appear to be fairly stable with the exception of the Bogardus Scale and the Trait Evaluation Scale on Self. Since the tests were administered two months apart, the effect of memory on the coefficient of reliability was probably insignificant (8).

As the relationship between the Bogardus Scale and other similar measures was of interest to the researcher, Pearson r's were computed between this scale and two adaptations of the Trait Evaluation Scale. The first adaptation of the scale explored the relationship between how a subject evaluated Negroes and his social acceptance of Negroes as measured by the Bogardus Scale. The second employment explored the relationship between 1) the differences between how a subject evaluated himself and how he evaluated the Negro, and 2) the Bogardus Scale. The results of these findings are listed in Table 2.

TABLE 2

CORRELATIONS BETWEEN BOGARDUS SCALE AND SIMILAR MEASURES ON THE NEGRO

Source	N	r_{xy}
Trait Evaluation Scale (Negro only)	82	.40*
Trait Evaluation Scales (Self-Negro)	82	.43*

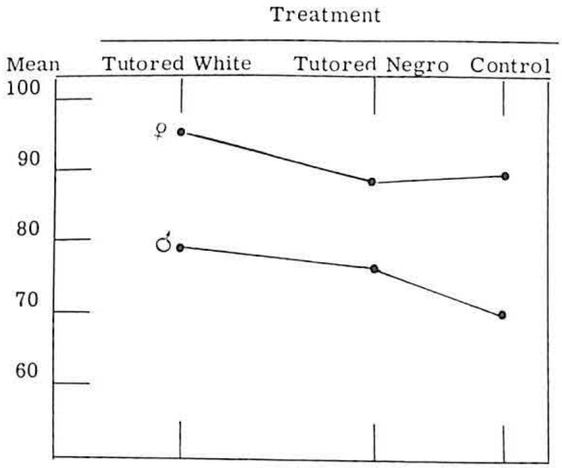
$^*P \leq .01$

The correlations between a subject's social acceptance of the Negro (Bogardus Scale) and his evaluation of him (Trait Evaluation Scale), both separately and with regard to himself, were surprisingly low (Table 2).

The statistical method of analysis of variance was used to estimate the degree to which change may have been involved in group differences. This procedure was followed as a method of testing the formal hypotheses. The first hypothesis questioned the effect of the treatment upon the subject's sensitivity to children. The MTAI was used to measure this variable. As it was not available for pre-testing, the analysis between groups could be made only on the post-test. Figure 1 illustrates the group means by sex and treatment.

FIGURE 1

GRAPHED MEANS OF MTAI BY SEX AND TREATMENT (N = 82)



Note: Constant of 50 added to all scores. High mean scores denote greater "sensitivity."

The results of the MTAI analysis of variance by sex and treatment reveal sex to be the only significant factor (Table 3). Female education students tend to score higher.

The second hypothesis was concerned with the effect of the helping relationship upon the subject's own feelings of personal and social adjustment. Two analysis of variance computations were undertaken. The first sought to discover whether, in fact, there was a difference between the experimental groups' pre- and post-tests. The performance means of the S-I Inventory by sex and time of testing are illustrated in Figure 2.

The analysis of variance did not reveal any significant differences due to the two independent variables (Table 4).

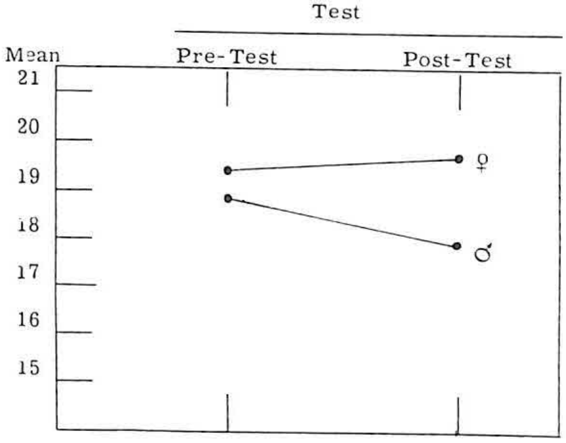
The second analysis of variance investigated performance differences on the S-I Inventory by sex and treatment. The means for this analysis are illustrated in Figure 3.

The performance scores for computing the analysis of variance of the S-I Inventory by sex and treatment were obtained by subtracting the post-test from the pre-test and adding a suitable constant to eliminate negatives. Computed differences were, therefore, differences in change.

The analysis of variance did not reveal any significant differences in change due to the two independent variables (Table 5).

FIGURE 2

GRAPHED MEANS OF S-I INVENTORY BY SEX AND TEST (N = 96)

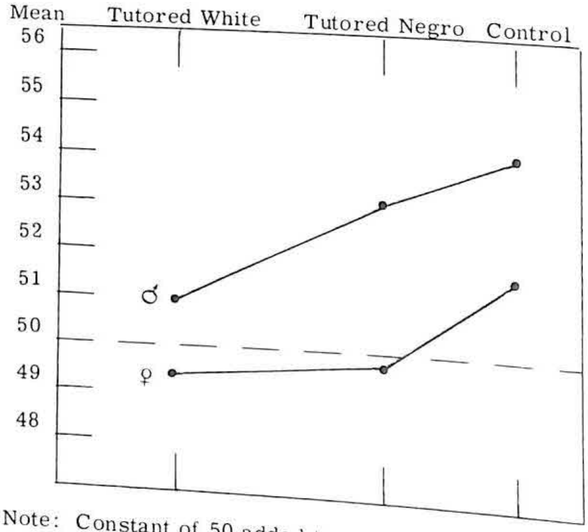


Note: Low scores denote greater "security."

The third hypothesis was concerned with the effect of the helping relationship upon the subject's cognitive flexibility as measured by the Dogmatism Scale. The procedure of analysis for the scale was identical with that employed for the S-I Inventory. Two analyses were undertaken. The performance means achieved on the Dogmatism Scale by sex and time of testing are illustrated in Figure 4.

FIGURE 3

GRAPHED MEANS OF S-I INVENTORY BY SEX AND TREATMENT (N = 82)



Note: Constant of 50 added to all difference scores $(X_1 - X_2)$. Mean scores above 50 denote change toward security.

TABLE 3

MTAI ANALYSIS OF VARIANCE BY SEX AND TREATMENT (N = 82)

Source	Sum of Squares	df	Variance Est.	F
Between Cells				
Experimental-Control (rows)	1,010. 57	2	505. 29	N. S.
Sex (columns)	4,716. 69	1	4,716. 69	5. 93*
Interaction	41. 17	2	20. 59	N. S.
Within Cells	60,169. 48	76	794. 75	
Total	66,169. 48	81		

*P ≤ .05

TABLE 4

S-I INVENTORY ANALYSIS OF VARIANCE BY SEX AND TEST (N = 96)

Source	Sum of Squares	df	Variance Est.	F
Between Cells				
Pre-Post Test (rows)	2. 34	1	2. 34	N. S.
Sex (columns)	19. 63	1	19. 63	N. S.
Interaction	17. 53	1	17. 53	N. S.
Within Cells	17,024. 99	92	185. 05	
Total	17,065. 49	95		

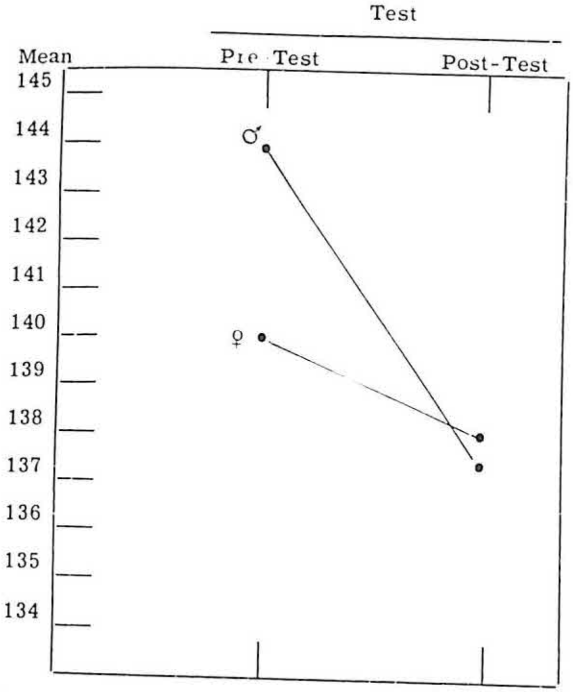
TABLE 5

S-I INVENTORY OF ANALYSIS OF VARIANCE BY SEX AND TREATMENT (N = 82)

Source	Sum of Squares	df	Variance Est.	F
Between Cells				
Treatment (rows)	100. 88	2	50. 44	N. S.
Sex (columns)	92. 20	1	92. 20	N. S.
Interaction	13. 81	2	6. 91	N. S.
Within Cells	2,643. 21	76	34. 78	
Total	2,850. 10	81		

FIGURE 4

GRAPHED MEANS OF DOGMATISM SCALE BY SEX AND TEST (N = 96)



Note: Low scores denote liberalism.

The analysis of variance for the Dogmatism Scale by sex and test did not reveal any significant differences due to the two independent variables (Table 6).

The second analysis of variance having to do with the dogmatism variable investigated performance differences by sex and treatment. The means for this analysis are illustrated in Figure 5.

The analysis of variance for the Dogmatism Scale by sex and treatment revealed no significant differences in change due to either of the independent variables (Table 7).

The fourth hypothesis explored the question of the effect of the helping relationship upon the subject's evaluation of the Negro in relation to his evaluation of himself. To accomplish this, each subject's evaluation of the Negro was subtracted from his trait evaluation of himself for both pre- and post-test differences. The result was a score representative of the degree of change in the way each subject evaluated the Negro relative to himself, i.e., $[(\text{Self}_1 - \text{Negro}_1) - (\text{Self}_2 - \text{Negro}_2)]$.

High scores indicated greater closure between the self-image and the Negro-image. The means of the converted scores from the Trait Evaluation Score are recorded in Figure 6 by sex and treatment.

The analysis of variance gave no indication of any significant effect upon image closure due to the treatment or sex (Table 8).

The final hypothesis of this study concerned the effect of "dogmatism" and "security" on group performance on the MTAI.

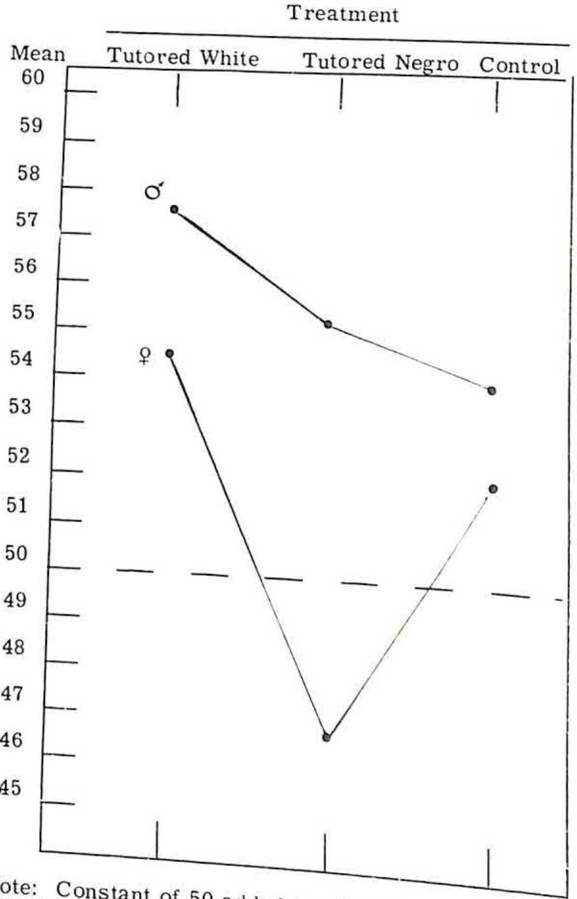
The entire sample of subjects was divided into three groups by their performance on the Dogmatism Scale and three groups by their performance on the S-I Inventory. The middle group in both cases consisted of those subjects within a half standard deviation of the sample mean.

The performance means of the MTAI by "dogmatism" and "security" are illustrated in Figure 7.

The MTAI analysis of variance by security and dogmatism revealed a very significant difference between dogmatic and liberal groups in their performance on the MTAI (Table 9).

FIGURE 5

GRAPHED MEANS OF DOGMATISM SCALE BY SEX AND TREATMENT (N = 82)



Note: Constant of 50 added to all difference scores $(X_1 - X_2)$. Mean scores above 50 denote change toward liberalism.

TABLE 6

DOGMATISM ANALYSIS OF VARIANCE BY SEX AND TEST (N = 96)

Source	Sum of Squares	df	Variance Est.	F
Between Cells				
Pre-Post Test (rows)	372.09	1	372.09	N.S.
Sex (columns)	70.20	1	70.20	N.S.
Interaction	105.39	1	105.39	N.S.
Within Cells	51,858.56	92	563.68	
Total	52,406.24	95		

TABLE 7

DOGMATISM ANALYSIS OF VARIANCE BY SEX AND TREATMENT (N = 82)

Source	Sum of Squares	df	Variance Est.	F
Between Cells				
Treatment (rows)	310.40	2	155.20	N.S.
Sex (columns)	251.83	1	251.83	N.S.
Interaction	132.71	2	66.36	N.S.
Within Cells	15,331.84	76	201.73	
Total	16,026.78	81		

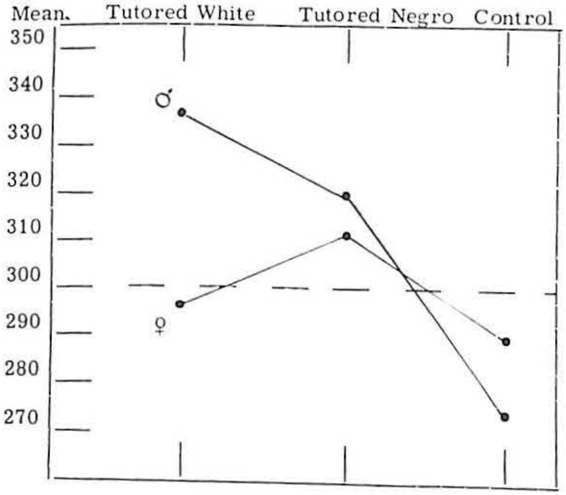
TABLE 8

ANALYSIS OF VARIANCE OF IMAGE CLOSURE BETWEEN SELF AND NEGRO BY SEX AND TREATMENT (N = 82)

Source	Sum of Squares	df	Variance Est.	F
Between Cells				
Treatment (rows)	19,793.65	2	9,896.83	N.S.
Sex (columns)	2,304.97	1	2,304.97	N.S.
Interaction	8,595.90	2	4,297.95	N.S.
Within Cells	391,678.41	76	5,153.66	
Total	422,372.93	81		

FIGURE 6

GRAPHED MEANS OF IMAGE CLOSURE BETWEEN SELF AND NEGRO BY SEX AND TREATMENT
(N = 82)



Note: Constant of 300 added to all difference scores $[(S_1-N_1) - (S_2-N_2)]$. Mean scores above 300 denote Image Closure.

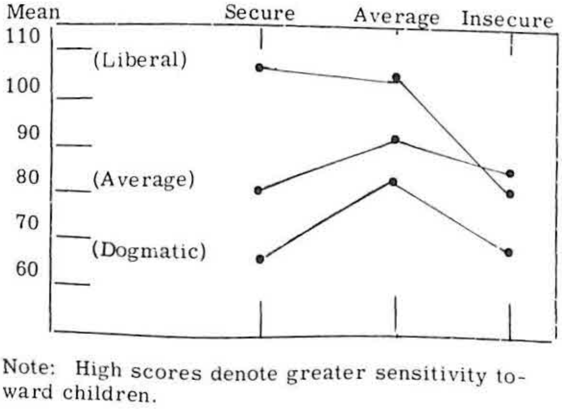
"Liberal" groups, as illustrated in Figure 7, tend to score higher on the MTAI than "dogmatic" groups. These differences appear to diminish as subjects move toward a state of insecurity. The differences due to extent of security, however, were not statistically significant.

CONCLUSION AND DISCUSSION

The statistical procedures for analyzing the data included correlation analysis and complex analysis of variance. All tests of significance were two-tailed. From the relatively high correlations achieved between the pre- and post-tests, it may be concluded that the stability of the instruments was adequate for the purpose of this study. The relatively low correlation between pre- and post-tests for the Trait Evaluation of Self in comparison with the correlation for the Trait Evaluation of the Negro was interesting. It may be that a person's beliefs about himself are subject to more variance than his beliefs concerning others. Whether this is only true of teacher education students with regard to Negroes is uncertain. It does, however, point up a significant degree of uncertainty and/or a striking amount of change in the self-concept at this stage.

FIGURE 7

GRAPHED MEANS OF MTAI BY DOGMATISM AND SECURITY (N = 82)



Note: High scores denote greater sensitivity toward children.

The second use of correlation analysis involved the Bogardus Social Distance Scale with another measure felt to be similar in nature. The obvious intent of the Bogardus Scale is to explore the degree to which a subject is willing to accept a pre-selected individual or type socially. The Trait Evaluation Scale was intended to reveal the subject's feelings for the individual, both independently and relative to the subject's own evaluation of himself. The low correlations indicated a poor relationship between the degree to which subjects are willing to accept another subject and the way they evaluate him as a person. Evidently, regard for an individual is not a clear indication of acceptance. It may be that the Trait Evaluation Scale fell short of specific areas related to acceptance. Another possibility is that responses on the Trait Evaluation Scale are simply not representative of what might be expected from the Bogardus Scale due to the subject's own unawareness of his true feelings. In such a case, the subject may be quite willing to recognize desirable qualities in another person but unable to accept him socially. Recognition of an individual as being similar to one's self, evidently, does not assure his social acceptance.

Only one of the five hypotheses was supported to any degree by this investigation. The treatment was apparently not of sufficient type or duration to significantly affect the variables under study. Although the differences in means were often in the anticipated direction, the variance within groups was of such magnitude that the results could just as easily have been attributed to chance.

The analysis of variance revealed only two significant F's, neither having to do with the "helping relationship." In the first instance, the female subjects scored significantly higher on the MTAI than the male subjects. To the extent that the MTAI is a valid measure of a subject's ability to communicate with children, it may be concluded that female education students, in most cases, will have a decided advantage in the classroom.

The other significant finding gave partial support to the last hypothesis. In seeking support for this hypothesis, the investigator was led to explore the effects of "dogmatism" and "security" on the subjects' MTAI performance. The findings indicated that a subject's general feeling of security has little effect upon his sensitivity to children. Cognitive flexibility, however, does appear to be a significant factor. "Liberal" subjects scored significantly

TABLE 9

MTAI ANALYSIS OF VARIANCE BY SECURITY AND DOGMATISM (N = 82)

Source	Sum of Squares	df	Variance Est.	F
Between Cells				
Security (rows)	3,033.26	2	1,516.63	N.S.
Dogmatism (columns)	9,558.55	2	4,779.28	6.50*
Interaction	1,106.87	4	2,767.21	N.S.
Within Cells	53,361.77	73	735.11	
Total	67,361.77	81		

*P ≤ .005

higher on the MTAI than "dogmatic" subjects. It may be that the variable needing the most attention in teacher education is that of "open-mindedness." Such an interpretation seems theoretically reasonable when one considers education to be a process of communication. The teacher as both "receiver" and "sender" of information must have some awareness of how the child perceives a problem. Such an awareness on the part of the teacher tends to insure the transmission of information relevant to the child's own level of awareness. An effective teacher must, therefore, have the ability to constantly readapt psychologically to the ever changing perceptions of the child. Such demands made of a "rigid" or "dogmatic" individual would most certaily result in anxiety and frustration.

The Morrison and Romoser factor analytic study of the MTAI represents evidence in support of this theory (7). Their findings indicate that students who performed poorly on the MTAI tended to be 1) traditional in their beliefs, 2) anxious, 3) less intelligent, and 4) male.

With regard to the present study, no significant differences were found between groups due to anxiety (S-I Inventory). Groups classified as either "female" or "liberal" did, however, score higher on the MTAI. It would seem, therefore, that the variable measured by the MTAI is closely allied to philosophical or cognitive beliefs concerning "liberalism." Why a subject responds "traditionally," however, on the MTAI may conceivably result from any number of causes. It is doubtful that the effect of traditionalism upon a subject's sensitivity to children would be the same irrespective of cause.

The implications of this study clearly indicate the need for additional research in the area. The study also revealed the inability of the type of curriculum experiences being offered education students to make a significant impact on particular personality and cognitive variables when such an experience is restricted to a single semester.

The fact that such variables are believed to be directly related to the establishment of good rapport lends additional weight to the findings. Future investigations into the question of how teacher education can increase its effect upon its students should not be limited to the area of modification alone, as demonstrated in this study, but should also consider the effects of duration and intensification.

FOOTNOTES

1. The writers wish to acknowledge the services and assistance of many persons who contributed to this study. They are especially appreciative of the help given to them by Mr. Don Mills, Dr. Elizabeth Parker, Miss Sharon Congleton, their secretary, and the participating students, all of Eastern Kentucky University.
2. The work presented or reported herein was performed pursuant to a grant from the U.S. Office of Education, Department of Health, Education, and Welfare.
3. This test was not available for pre-testing.

REFERENCES

1. Asbell, B., "Let the Children Teach," *Redbook*, CXXIV (February, 1966), p. 53.
2. ASCD "Children Teach Each Other," *ASCD Yearbook* (1962), pp. 32-35.
3. Gill, N. T., *An Analysis of Student Reaction to Selected Treatments Used in Education 316*, an unpublished study, Eastern Kentucky University, 1966.
4. Howe, H. II. *A New Benchmark for Education*, an address by Harold Howe II, U.S. Commissioner of Education; Department of Health, Education, and Welfare, 1966.
5. Karowe, H. E., "How Volunteers Can Help Disadvantaged Children," *Children*, XIV (1967), pp. 151-155.

6. Krech, D., Crutchfield, R. S., and Ballachey, E. L., Individual in Society (New York: McGraw-Hill, 1962).

7. Morrison, W. L., and Romoser, R. C., "Traditional Classroom Attitudes, the A. C. T. and the 16 P. F.," Journal of Educational Research, LX (1967), pp. 326-329.

8. McNemar, A., Psychological Statistics (New York: Wiley, 1963).

9. Rogers, C. R., On Becoming a Person (Boston: Houghton-Mifflin, 1961).

10. Rokeach, M., The Open and Closed Mind (New York: Basic Books, 1960).

APPENDIX A

Social Distance Scale

Date ____________ Name ____________

Score ____________

Type of Individual Being Considered ____________

The instructions for the scale are as follows:

"According to my first feeling reactions, I would willingly admit the type of individual considered on the S-D Scale No. 1 to one or more of the classifications under which I have placed a cross."

____ 1. To close kinship by marriage.
____ 2. To my club as a personal chum.
____ 3. To my street as a neighbor.
____ 4. To equal employment in my occupation.
____ 5. To citizenship in my country.
____ 6. As a visitor in my country.
____ 7. To citizenship in a foreign country.
____ 8. As a member of the human race.

APPENDIX B

Trait Evaluation Scale

Date ____________ Name ____________ Score ____________

Individual Being Considered ____________

			Score
1. Progressive	1 2 3 4 5 6 7	Conservative	
2. Honest	1 2 3 4 5 6 7	Dishonest	
3. Physically clean	1 2 3 4 5 6 7	Physically dirty	
4. Respectable	1 2 3 4 5 6 7	Indecent	
5. Reliable	1 2 3 4 5 6 7	Unreliable	
6. Compassionate, warm-hearted	1 2 3 4 5 6 7	Unforgiving, unmerciful	
7. Intelligent	1 2 3 4 5 6 7	Unaware, ignorant	
8. Loyal	1 2 3 4 5 6 7	Unfaithful	
9. Industrious, hard-working	1 2 3 4 5 6 7	Lazy, indolent	
10. Practical, sensible	1 2 3 4 5 6 7	Foolish, irrational	
11. Creative, enterprising	1 2 3 4 5 6 7	Unimaginative, traditional	
12. Alert	1 2 3 4 5 6 7	Dull, slow	
13. Courteous, polite	1 2 3 4 5 6 7	Bad-mannered, boorish, rude	
14. Humorous	1 2 3 4 5 6 7	Critical, solemn, grim	
15. Ambitious, eager	1 2 3 4 5 6 7	Unmotivated, indifferent	
16. Persistent, persevering	1 2 3 4 5 6 7	Yielding, submissive	
17. Friendly	1 2 3 4 5 6 7	Hostile, quarrelsome	
18. Altruistic (likes people)	1 2 3 4 5 6 7	Misanthropic (hates people)	
19. Productive	1 2 3 4 5 6 7	Useless, ineffective	
20. Likeable, enjoyable	1 2 3 4 5 6 7	Obnoxious, offensive	
21. Scientific	1 2 3 4 5 6 7	Superstitious	
22. Attractive	1 2 3 4 5 6 7	Repulsive	
23. Optimistic	1 2 3 4 5 6 7	Pessimistic	
24. Realist	1 2 3 4 5 6 7	Idealist	

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

KINDERGARTEN TEACHER AIDES AND READING READINESS MINNEAPOLIS PUBLIC SCHOOLS

PATRICIA J. GORALSKI and JOYCE M. KERL

ABSTRACT

In order to discover whether teacher aides can be used effectively to help develop reading readiness in kindergarten children, three groups of three classes each were compared: (1) Group A, no teacher aide, (2) Group B, one teacher aide, (3) Group C, five teacher aides. Under the direction of the teachers, aides worked with children in activities designed to increase reading readiness. There were no initial differences between groups (Metropolitan Readiness Test, Form R) , but there were significant differences at the end of one semester in favor of the groups with aides.

TEACHER AIDES have served the Minneapolis Public Schools in increasing numbers since the summer of 1965. During the school year, 1966-67, 260 teacher aides worked in Minneapolis target area schools. Periodic evaluations have indicated that their work has been well received by both educators and the community. The aides have been enthusiastic about their participation in the educational experiences of children.

PURPOSE

The first experimental attempt to discover a relationship between aides' services and pupil gain was made during spring semester, 1967. Nine classes in four Minneapolis target area schools participated in an experiment which attempted to answer the question: "Can teacher aides be used effectively to help to develop reading readiness in kindergarten children?"

SAMPLE SELECTION CRITERIA

Nine kindergarten classes taught by six teachers were identified which met the following criteria:

--Each class consisted of approximately thirty children.

--Approximately half of the classes met in the morning and half in the afternoon.

--None of the classes had had the services of aides during the fall semester, 1966.

--Only classes from target area schools that did not have lunch programs were included. (Where lunch programs existed, afternoon kindergarten sessions were shorter).

--Of the thirty children in each classroom, approximately one-third had participated in Headstart during the summer of 1966. (The reason for this criterion was to furnish data for long-range Headstart evaluation).

--Three of the teachers were to have both morning and afternoon classes which met the other criteria.

DESIGN AND TESTING PROCEDURE

The null hypothesis was:

> H_0: There is no difference between groups of second semester kindergarten pupils in reading readiness mean gain scores on the Metropolitan Readiness Test.

A diagram of the design of the experiment has been presented in Table 1.

Since the researchers wanted the strongest control for teacher effects for the group with no teacher aides and the group with five teacher aides, six of the classes constituted the morning and afternoon classes of three teachers and made up Groups A and C. Each of these teachers had no aide in one class and five aides in the other. Teachers of Groups A

TABLE 1

THE EXPERIMENTAL DESIGN

Group A No teacher aide	Group B One teacher aide	Group C Five teacher aides
Class 1, Teacher 1 School W, P. M.	Class 4, Teacher 4 School Y, A. M.	Class 7, Teacher 1 School W, A. M.
Class 2, Teacher 2 School X, A. M.	Class 5, Teacher 5 School Y, P. M.	Class 8, Teacher 2 School X, P. M.
Class 3, Teacher 3 School X, A. M.	Class 6, Teacher 6 School Z, A. M.	Class 9, Teacher 3 School X, P. M.
N=38 M N=37 F N=75 Total, Group A	N=39 M N=40 F N=79 Total, Group B	N=43 M N=37 F N=80 Total, Group C

and C were careful to use the added assistance for Group C classes, only, so that effects would not be contaminated. Group B consisted of the classes of three different teachers with one aide assigned to each class.

Further, teacher effects tended to cancel because only group effects were analyzed and each group consisted of three different teachers. However, the strongest control for teacher effects was provided in that

--pretests were not administered until the end of the first semester when teachers had already worked with their classes for half the year

--there were no initial differences between groups nor was there interaction

--the same teachers continued with their classes throughout the second semester.

Reading readiness was assessed by the Metropolitan Readiness Test, Form R, (1949). Pretests were administered to 254 children in the nine classes during the period from January 25, 1967, through February 3, 1967. Posttests were administered to 248 children during the period from May 22, 1967, through May 26, 1967. Testing yielded 234 pairs of test scores. Aides served six of the nine classrooms from February 6, 1967, through June 2, 1967.

TREATMENT AND RATIONALE

It was hypothesized that a teacher would be likely to view one aide as an extension of herself, as an assistant to help her with whatever she was doing. It would not be necessary for the teacher with one aide to develop a concept of the role of the aide separate from her own activity.

It was further hypothesized that it would be impossible for teachers with five aides in one classroom to proceed without giving a great deal of thought to the kinds of things that aides might accomplish which would be in addition to, or different from, what she might accomplish alone. The design of the experiment made it necessary for the three teachers of Groups A and C to think about differences in the approach to problems of teaching and learning where there was only the teacher in the classroom (Group A) and where there were five adults in addition to the teacher (Group C).

During two inservice meetings, teachers exchanged ideas about the effective use of aides. Teachers were encouraged to use whatever approaches they thought might be effective in achieving maximum reading readiness gains with pupils. It was emphasized that, with more adults in the classroom, there could be almost total participation of all children at all times: More physical activity, more verbal expression, more interaction. The tendency of adults to do things for children rather than helping children to learn to do things for themselves was explored. However, each teacher was completely free to use the services of her aides in any way that she believed to be best.

It was an unsettling experience for a teacher to add five adults to one classroom and classroom administration problems arose. However, as the semester progressed, teachers discovered that there were important services that five adults could perform in one classroom. Teachers expressed an interest in trying a similar experiment in the future. One of the teachers stated that she would like to have all five of her aides again the following year.

It would have been valuable and interesting if the three teachers of Groups A and C had kept a diary to record:

--their feelings as five adults came into one kindergarten classroom

--the problems that arose, both among adults and with children

--techniques used to solve problems

- -*ways that they tried to use aides' services effectively to increase reading readiness in children*
- --areas where aides were successful in enhancing pupil learning
- --areas where aides did not seem to be successful in helping pupils to advance their learning.

However, since the addition of five adults to one kindergarten classroom was a rather overwhelming prospect for any teacher, it was feared that added paperwork might change very interested and cooperative teachers into rather reluctant ones. It is recommended that future experimentation include provision for participants to record their feelings about working with adults in the classroom as well as their attempts to work through teaching and learning problems. A weekly tape-recorded meeting would be an effective technique for ascertaining both essence and fact.

ANALYSIS AND INTERPRETATION OF THE DATA

In order to make a meaningful study of the major hypothesis, it was necessary to find out whether, at the beginning of the study, groups differed with respect to reading readiness characteristics. Since it has been widely held that there are differences in readiness by sex, the following sub hypothesis was tested.

> H_1: There is no difference in pretest mean scores on the Metropolitan Readiness Test between groups or by sex.

Three analyses of variance were calculated, one for each of three aspects of the test. They were reported in Tables 2, 3, and 4.

TABLE 2

ANALYSIS OF VARIANCE, PRETEST SCORES ON THE METROPOLITAN READINESS TEST, FORM R TESTS 1-4, READING READINESS

Source of Variation	df	SS	MS	F
Groups	2	14.22	7.11	1.71
Sex	1	16.80	16.80	4.05
Interaction	2	4.58	2.29	0.55
Error, Within Cells	228	36804.75	4.15,adjusted	

Critical Value: $F_{.95}\ (1, 228) = 3.8851$

$F_{.95}\ (2, 228) = 3.0367$

TABLE 3

ANALYSIS OF VARIANCE, PRETEST SCORES ON THE METROPOLOTAN READINESS TEST, FORM R TEST 5, NUMBER READINESS

Source of Variation	df	SS	MS	F
Groups	2	0.18	0.09	0.15
Sex	1	1.59	1.59	2.56
Interaction	2	1.63	1.31	2.11
Error, Within Cells	228	5513.41	0.62 adjusted	

Critical Value: $F_{.95}\ (1, 228) = 3.8851$

$F_{.95}\ (2, 228) = 3.0367$

TABLE 4

ANALYSIS OF VARIANCE, PRETEST SCORES ON THE METROPOLITAN READINESS TEST, FORM R TEST 1-6, TOTAL READINESS

Source of Variation	df	SS	MS	F
Groups	2	17.41	8.71	0.99
Sex	1	34.71	34.71	3.96
Interaction	2	15.43	7.72	0.88
Error, within cells	228	77663.96	8.76, Adjusted	

Critical Value: $F_{.95}\ (1, 228) = 3.8851$

$F_{.95}\ (2, 228) = 3.0367$

It was found that there were no significant differences, by groups, at the 5 percent level. However, there were significant differences at the 5 percent level of confidence, by sex, in reading readiness and total readiness. Therefore, analysis of covariance, which allowed for initial differences, was used to test the major hypothesis:

> H_0: There is no difference between groups of second semester kindergarten pupils in reading readiness mean gain scores on the Metropolitan Readiness Test.

Three analyses of covariance were calculated and reported in Tables 5, 6, and 7.

It was found that, after initial differences had been adjusted by means of analysis of covariance, there were no significant differences, by sex, at the 5 percent level in reading readiness, numbers readiness, or total readiness. It was concluded that whatever differences there were by sex, they were not related to the treatment (the services performed by teacher aides).

TABLE 5

ANALYSIS OF COVARIANCE, PRETEST AND POSTTEST SCORES ON THE METROPOLITAN READINESS TEST, FORM R TEST 1-4, READING READINESS

Source of Variation	df	SS	MS	F
Groups	2	9.14	4.57	3.83
Sex	1	1.40	1.40	1.17
Interaction	2	1.36	0.68	0.57
Error, within cells	227	270.78	1.19	Adjusted

Critical Value: $F_{.95}(1, 227) = 3.8859$
$F_{.95}(2, 227) = 3.0373$

TABLE 6

ANALYSIS OF COVARIANCE, PRETEST AND POSTTEST SCORES ON THE METROPOLITAN READINESS TEST, FORM R TEST 5, NUMBER READINESS

Source of Variation	df	SS	MS	F
Groups	2	1.12	0.60	1.59
Sex	1	0.90	0.90	2.55
Interaction	2	0.24	0.12	0.35
Error, within cells	227	79.94	0.35,	Adjusted

Critical Value: $F_{.95}(1, 227) = 3.8859$
$F_{.95}(2, 227) = 3.0373$

TABLE 7

ANALYSIS OF COVARIANCE, PRETEST AND POSTTEST SCORES ON THE METROPOLITAN READINESS TEST, FORM R TESTS 1-6, TOTAL READINESS

Source of Variation	df	SS	MS	F
Groups	2	17.27	8.63	4.34
Sex	1	5.82	5.82	2.92
Interaction	2	4.51	2.26	1.13
Error, within cells	227	451.72	1.99,	Adjusted

Critical Value: $F_{.95}(1, 227) = 3.8859$
$F_{.95}(2, 227) = 3.0373$

The major findings, based on analysis of covariance were that there were significant differences between groups at the 5 percent level in reading readiness (tests 1-4) and total readiness (tests 1-6). Analysis of covariance revealed no differences between groups, significant at the 5 percent level in numbers readiness (test 5). Therefore, no further discussion of numbers readiness has been included.

In order to study the magnitude of gains for each group, a table of unweighted cell means has been presented (Table 8).

TABLE 8

UNWEIGHTED CELL MEANS, PRETEST, POSTTEST, AND GAIN METROPOLITAN READINESS TEST, FORM R (N = 234)

Test	Group A N=75	Group B N=79	Group C N=80
Tests 1-4, Reading Readiness			
Pretest	41.8	38.2	40.7
Posttest	47.4	47.8	48.8
Gains	5.6	9.6	8.1
Test 5, Number Readiness			
Pretest	8.2	8.1	8.4
Posttest	11.6	12.1	12.7
Gains	3.4	4.0	4.4
Tests 1-6, Total Readiness			
Pretest	53.5	49.7	52.8
Posttest	63.9	64.8	66.4
Gains	10.4	15.2	13.6

An inspection of Table 8, revealed that Group B classes with a teacher and one teacher aide, experienced the greatest mean gains on tests of reading readiness and total readiness. Group A, the classes with a teacher and no aides, experienced the least mean gains in reading readiness. Group C mean gain scores fell between those of Group A and Group B.

Scheffé's test was used to analyze between groups differences for total readiness, tests 1-6. It was found that differences between Group B and Group A were significant at the 5 percent level. Differences between Groups C and Group A approached this level. Differences between Group B and Group C did not approach significance at the 5 percent level.

Scheffé's test was also used to analyze between groups differences for reading readiness, tests 1-4. It was found that none of the differences between groups were significant at the 5 percent level. However, differences between Group B and Group A, and differences between Group C and Group A, approached significance at the 5 percent level. Differences between Groups B and C did not approach significance.

It was concluded that teacher aides can be used effectively to help to develop reading readiness in kindergarten children. Children had gained more in reading readiness when aides were present in the classroom.

It was found that children made slightly greater gains in the group where only one aide worked in the classroom than in the group with five aides. However, it did not seem to make much difference whether there were five aides in the room, or one. It was inferred that when five aides were present in the classroom, the teacher spent more time training and supervising aides and less time working directly with children than she did when only one aide was present. Since teachers are professionally trained to work with children, it was believed that directing the teacher's efforts elsewhere might not result in optimum pupil gain. Perhaps, if aides had been trained before coming into the classroom to work with children toward a specific classroom objective, such as increasing reading readiness, then aides' classroom activities might not have required the degree of teacher planning and supervision that was necessary during the experiment.

Results of this experimental study have indicated that the services of teacher aides can help to develop reading readiness in kindergarten children. It is recommended that future research programs investigate:

--the kinds of training necessary to make aides most effective in areas where their work has been found to result in pupil gains.

--the numbers of aides necessary to achieve optimum pupil gains in areas where it has been found that aides' services result in pupil gains.

--other areas where the services of aides result in significant pupil gains.

The researchers believe that the services of trained aides should free the teacher to work more closely with children in the areas where greatest professonal skill and competence are required. It is recommended that aides continue to work directly with children in areas where it is discovered that their services are most effective. However, whatever the number of aides present in the classroom and whatever the kinds of work performed, aides' services should enable the teacher to spend more time with children rather than less.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

TEACHER ACCEPTANCE OF ADMINISTRATIVE ACTION

JOY GOLD HARALICK*
University of Kansas

ABSTRACT

Democratic administrative leadership has long been hypothesized to result in high worker morale. Within the context of the elementary school, the principal is the chief administrator and the teachers occupy the major adult work roles. Data from 568 teachers in 52 elementary schools, classified as to administrative atmosphere, indicates that the principal's compliance with specific work-related norms held by the teacher is more important to teacher satisfaction than is the degree of democratic behavior which he exhibits. Some explanations for this phenomenon are suggested.

BACKGROUND

The human relations literature tends to divide administrative leadership types into two broad categories: "democratic" and "autocratic." These categories have also been called by other names such as "member-centered" and "leader-centered," "non-authoritarian" and "authoritarian," "non-directive" and "directive," "lenient" and "strict." The "democratic" leader is usually defined operationally as one who allows the group members some participation in decision making processes; who is available to give and to receive suggestions and advice; and who gives praise based on standards of performance which are known by the group members. Conversely, the "autocratic" leader is operationally defined as one who makes decisions without consulting the group members; who gives advice but is unwilling to receive it; and who gives arbitrary praise and criticism based on standards which are unknown to group members. According to much of the research in this area, groups with "democratic" leaders tend to exhibit higher morale, a greater degree of job satisfaction, fewer cliques, less hostility among members, greater approval of their leaders, and a higher degree of worker effectiveness. A third general type of leadership pattern has been designated by Lippitt and White as laissez-faire. The laissez-faire leader plays a passive role in social participation and gives the group members complete freedom in decision making. This type of leader neither evaluates nor takes the initiative in making suggestions. The present study omits laissez-faire leadership from consideration, because the nature of most formal organizations limits the amount of goal-defining which may be done by subordinates in the hierarchical structure, and requires evaluation of their work role performances by their administrative superiors. It should be pointed out that the laissez-faire leadership type seems to be truly democratic in the "people's rule" sense of the term, whereas what is usually designated as "democratic" leadership is only democratic relative to "autocratic" leadership, and may be more precisely labeled "lenient," "participant-oriented," etc. Regardless of the label, according to most of the human relations studies, "democratic" adminstration seems to constitute a very effective type of leadership behavior.[1]

Although the human relations model tends to be a bureaucratic one,[2] there are, of course, many different types of bureaucratic organizations. Etzioni categorizes formal organizations in terms of the major types of power used. Utilitarian organizations are those which utilize remunerative power or "control over material resources and rewards." Normative organizations use normative power or the "allocation and manipulation of symbolic rewards and deprivations." Coercive organizations use coercive power or the threat or actual application of physical sanctions. An example of a utilitarian organization is an industry; of a coercive organization, a prison; of a normative organization, a hospital.[3] Most human relations studies have been concerned with utilitarian organizations. Even where the organizations have been normative ones, the supervisor-subordinate relationships studied have tended to be those in which coercion or remuneration play a dominant role.[4]

The present study focuses on the relationship between the two major normative roles – one administrative and one subordinate – in a predominantly normative organization. The roles are those of teacher and principal. The organization is the public elementary school. The study examines the applicability of current human relations theory to the school as a normative, bureaucratic institution.

Teachers, besides being members of normative organizations, differ in other respects from the industrial workers who have been the subjects of most human relations research. They tend, at the elementary school level at least, to be women rather than men. As women, given the values of our society, they may be expected to plan their careers on the bases of factors extrinsic to the work situation. These factors may included such life circumstances as marriage or spinsterhood, wealth or poverty, children or absence of children.[5] Teachers also tend to have had more education than most of the assembly line workers studied.[6] Their jobs are less routine than the assembly line job. The teacher has more decisions to make and is less closely supervised than the industrial worker. For all these reasons: sex, education, autonomy, and lack of routine, the possibility exists that the teacher may respond to certain types of administrative behavior in a somewhat different manner than would an industrial worker. Still, the teacher is a worker in a formal organization, and her work is subject to orders and evaluation from the principal who is her hierarchical superior. She should then, if the human relations studies are correct, respond to certain general adminstrative practices in the same direction, if not with the same intensity as industrial workers. The major questions with which this paper is concerned are the effects of particular types of adminstrative atmosphere on the teacher's evaluation of the principal's leadership ability; on her work satisfaction; on the amount of cohesiveness the teachers believe exists among themselves; on the perception of the existence of hierarchical cliques; on the attitudes of the teachers toward the pupils and their parents; and on the teacher's feelings of freedom to consult the principal. These are the variables which in most human relations studies are believed to be dependent on the degree of "democratic" or "autocratic" behavior exhibited by the leader or supervisor of the work group.[7]

HYPOTHESIS

Merton suggests that the effective supervisor must know the norms of his subordinates. Otherwise he may unintentionally violate these norms, and the violations may cause resentment which eventually leads to the undermining of the leader's authority and to the inefficient pursuit of organizational goals.[8] Since subordinate participation and effective two way communication have been the main variables used to operationally define "democratic" leadership, it seems quite possible that from the administrator's position, "democratic" leadership functions as a means to the end of getting to know subordinate norms. Once these norms are known, the adminstrator will be able to then to use this knowledge to ingratiate himself with the subordinates, thus making them more willing to do good work for him and for the organization of which he is a representative. This is not necessarily a conscious process.[9] Related to this is the implication that in any organization there are other norms, administrative compliance with which, from the subordinate's viewpoint, is more important than administrative adherence to "democratic" behavior. In other words, "democratic" behavior is a means to the end of learning about the important norms which the subordinates hold.

It has been shown in several studies that teachers are most satisfied with their jobs when the principal and the teachers are in agreement on occupational norms. It has been shown in these same studies that teachers hold clear norms concerning what constitutes acceptable behavior by the principal. The most important norms which teachers hold concerning the principal's behavior are:

1. He should "back up" (i. e., support) the teacher in front of parents, even when he considers the teacher to be "wrong." If he criticizes her, it should be done privately.
2. He should "back up" the teacher in front of pupils. Again, if the principal believes that the student is "right," the teacher should be told this privately. She should never be criticized in the presence of the pupil.
3. He should be a good disciplinarian.
4. He should exhibit democratic behavior.[10]

Although these are all norms with which the teacher expects the principal to comply, we would expect if "democratic" behavior is primarily useful as a means to the knowledge of subordinate norms, administrative compliance with norms 1 through 3 to be more important in determining aspects of teacher satisfaction than compliance with norm 4. As a form of shorthand throughout this paper, the norms of "backing" and disciplinarianship are referred to as "ends norms;" administrative compliance with the ends norms is called "positive normative compliance;" non-compliance with the ends norms is called "negative normative compliance." The administrator (the principal in this case) can exhibit positive compliance with the ends norms and yet be "autocratic," or he can exhibit negative compliance with the ends norms and be "democratic," or he can comply or refuse to comply with both sets of norms. This combination of possibilities yields a typology of four kinds of administrative atmosphere: democratic-positive (compliance with ends norms), autocratic-positive, democratic-negative, and autocratic-negative.

The basic hypothesis of this study is that for any index of teacher satisfaction traditionally thought to be positively influenced by "democratic" leadership, a stronger positive influence will be exerted by perception of the principal's positive compliance with the ends norms. The highest percentage of most satisfied teachers should be in "democratic" schools where the principal shows positive compliance with the ends norms. The next highest percentage of most satisfied teachers should be found in "autocratic" schools in which the principal exhibits positive normative compliance. The percentage of most satisfied teachers on any given index of satisfaction should decline further in "democratic" schools where the principal is non-compliant and should reach its low point in autocratic schools with non-compliant principals. Table 1 provides a graphic illustration of this hypothesis, as well as an example of the type of percentage table by means of which the hypothesis is examined.

THE SAMPLE

The data for this study are obtained from an analysis of the questionnaire responses of 1,250 white teachers in 108 North Carolina elementary schools.[11] All of the subjects come from schools containing 14 to 25 teachers, the modal size category for elementary schools in the State. Due to the method of using teacher consensus to classify schools according autocracy of administrative atmosphere (see next section for further explanation), teachers from schools with responses rates of under 50 percent are omitted from the sample. This restriction, the restriction on size of school, and another one which eliminates from the sample teachers who are also part of administration, lowers the sample number from the 1,586 (49.3 percent of the white elementary school teachers in the State in the spring of 1962, when the data were collected) teachers who answered the questionnaire to 1,250 whose questionnaires are usable. Ninety-four and three-tenths percent of the sample is female. Most of the teachers have more than 30 pupils in a classroom. Over half of the teachers are over forty years of age; about one quarter of them are under thirty.[12]

MEASUREMENT

In order to test the hypothesis, each of the 108 elementary schools in the sample (it will be recalled that each of the schools has a response rate of over 50 percent) is classified according to autocracy of administrative atmosphere. The classifications are based on the response consensus of the teachers in each school to a six item Likert scale concerning the behavior of the principal. Each teacher has a score which is based on her responses to the scale items. The mean score is computed from the individual scores of the responding teachers in each school, thus giving each school an administrative atmosphere score, based on the collective responses of the teachers. The schools are grouped into three categories: low autocratic or "democratic" schools (314 teachers), no consensus (682 teachers) and highly autocratic schools (254 teachers). For the purposes of examining the hypothesis, only the schools in which there is consensus concerning the autocracy or lack of autocracy of the principal are included in the study. This additional restriction reduces the number of schools analyzed from 108 to 52 and the number of teachers from 1,250 to 568. The Likert items on which the teachers were asked to rate their principals are as follows:

1. Allowing teachers freedom on the job
   1.... Outstanding
   2.... Above Average
   3.... Average
   4.... Below Average
2. Letting teachers participate in decisions that concern them (same categories as 1, above)
3. Identifying himself and his interests with the teacher group (same categories as 1, above)
4. Supporting and encouraging teachers in new ways of doing things (same categories as 1, above).
5. Does your principal get together or talk informally with teachers or does he confine himself strictly to business?
   1.... Often associates informally with teachers
   2.... Associates informally with teachers from time to time
   3.... Rarely associates informally with the teachers
   4.... Never associates informally with the teachers
6. How often does the principal ask you for advice or for your opinion on school matters affecting you?
   0.... Less than once a month
   1.... 1 to 3 times a month
   2.... 4 to 6 times a month
   3.... 7 to 9 times a month
   4.... 10 times a month or more

A *t* test was used to examine the discriminatory power of each item. It was found that each item discriminated between the high and low quartile of autocracy scores at a probability level of less than .001.[13] A high score on the scale indicates a perception of a highly autocratic principal.

A simple index of normative compliance (compliance with the ends norms) was constructed from the following three items:

1. Can the principal be counted on to stand up for the teacher when a parent comes in with a complaint?
   1.... Can always be counted on
   2.... Can usually be counted on
   3.... Can sometimes be counted on
   4.... Can hardly ever be counted on
   5.... Can never be counted on
2. Can the principal be counted on to stand up for the teacher when a student comes in with a complaint?
   1.... Can always be counted on
   2.... Can usually be counted on
   3.... Can sometimes be counted on
   4.... Can hardly ever be counted on
   5.... Can never be counted on
3. Is your principal a good disciplinarian?
   1.... Outstanding
   2.... Above Average
   3.... Average
   4.... Below Average

The items are scored according to their code values. In items 1 and 2, codes 4 and 5 are combined because of small response numbers into one category, "4." Scores of 3, 4, or 5 are called "positive normative compliance." Scores of 6 or above are called "negative normative compliance."[14]

FINDINGS

According to the literature, many varied aspects of the work situation, especially those aspects closely pertaining to work satisfaction, tend to be influenced by administrative leadership behavior. On the basis of face validity and item intercorrelations, instruments were designed to measure some of these aspects. It will be recalled that it was hypothesized that for any index of teacher satisfaction, traditionally thought to be positively influenced by "democratic" leadership, a stronger positive influence will be exerted by perception of the principal's positive compliance with the ends norms of "backing" and disciplinarianship (positive normative compliance). The hypothesis is examined by means of a series of percentage tables, similar to Table 1. Because of the non-probability nature of the sample, statistical tests of significance are not used.

Human relations studies have generally found a positive relationship between "democratic" leadership and subordinate evaluation of the supervisor's leadership ability. In the present study, teachers were asked to rate the "overall leadership ability" of their principal. They were given four response alternatives ranging from "outstanding" to "below average." High leadership evaluation (see Table 2) is found to be more closely related to the perception

TABLE 1

PREDICTED PERCENTAGES OF TEACHERS EXHIBITING HIGHEST SCORES ON ANY INDEX OF SATISFACTION RELATED TO THE WORK SITUATION

	Administrative Atmosphere	
Normative Compliance	Democratic	Autocratic
Positive	highest percentage	next highest percentage
Negative	next lowest percentage	lowest percentage

TABLE 2

PERCENTAGE OF TEACHERS ACCORDING HIGHEST LEADERSHIP EVALUATION TO THE PRINCIPAL ("OUTSTANDING" RESPONSE)

	Democratic School		Autocratic School		
Normative Compliance	Percent	N^a	Percent	N^a	Total
Positive	49.6	252	19.5	128	380
Negative	9.4	53	0.0	117	170

N^a = the number of teachers who answered the questions in each type of school – the bases from which the percentages were computed. For example, 49.6 percent of the 252 teachers in democratic-positive schools (democratic schools where the principal complies with the "ends" norms) grant their principal an evaluation of outstanding leadership ability. None of the 117 teachers in autocratic-negative schools gave an "outstanding" leadership rating to their principal. All of the five tables which follow are alike in regard to the way in which they should be read.

of the principal's positive compliance with the ends norms than to the autocracy or democracy of his behavior.

"Democratic" leadership has been related to work satisfaction. An index of work satisfaction was constructed from two items which asked the teacher about her satisfaction with her specific job and with her teaching career. The items are:

1. How satisfied are you with your present teaching position, compared with what you imagine most teaching positions are like?
   1.... It is the best teaching position I can realistically imagine.
   2.... It is better than most teaching positions.
   3.... It is about as good as most teaching positions.
   4.... It is inferior to most teaching positions.
2. Speaking for yourself, how satisfying has teaching been as a career?
   1.... The most satisfying career I can imagine
   2.... More satisfying than most careers
   3.... About as satisfying as most careers
   4.... Less satisfying than most careers.

Scores on the index can run from 2 to 8 in the direction of most to least favorable. Highest work satisfaction (a score of 2) is more closely related to the perception of the principal's compliance with the ends norms than to his "democratic" behavior. (See Table 3.)

People who work in organizations in which there is "democratic" leadership have been shown by much human relations research to have higher morale. Because the teacher works so closely with the organizations's "clients," her opinion of the "clients" should be one index of her morale. Teachers were asked to "rate the attitude of the students in your school" and to "rate the attitude of the parents." [15] The four response alternatives range from outstanding to below average, a score of 2 indicating the highest rating on this two item index. Again the tendency appears for highest rating of clients to relate more strongly to the positive compliance of the principal with the ends norms than to his "democratic" behavior. (See Table 4.)

An index measuring feelings of cohesiveness

TABLE 3

PERCENTAGE OF TEACHERS RECEIVING HIGHEST SCORE ON WORK SATISFACTION INDEX (SCORE OF 2)

	Democratic School		Autocratic School		
Normative Compliance	Percent	N^a	Percent	N^a	Total
Positive	25.1	247	11.1	126	373
Negative	9.4	53	4.4	114	167

N^a = the number of teachers answering the questions in each type of school – the bases from which the percentages were computed.

TABLE 4

PERCENTAGE OF TEACHERS GIVING HIGHEST RATING TO "SCHOOL CLIENTS"

	Democratic Schools		Autocratic Schools		
Normative Compliance	Percent	N^a	Percent	N^a	Total
Positive	10.5	248	4.0	126	374
Negative	3.8	52	1.7	117	169

N^a = the number of teachers answering the questions in each type of school – the bases from which the percentages were computed.

with other teachers was constructed from two items which ask the respondent how well she likes the other teachers in the school and how similar she feels to them. The questions are:

1. How well do you like most of the other teachers in your school?
   1.... Very Well
   2.... Fairly Well
   3.... Not Much
   4.... Not at All
2. How similar are you to other teachers in your school in your values and outlook on life?
   1.... Similar to Most
   2.... Similar to About Half
   3.... Similar to Fewer Than Half

The index scores can run from 2 to 7 in the direction of most to least favorable. The highest cohesiveness score is a score of 2. Table 5 shows that the hypothesis that cohesiveness has a greater relationship to compliance with ends norms than to "democratic" behavior is upheld in the case of positive compliance but is not supported in the case of non-compliance.

Clique formation has been related to autocratic leadership in some of the human relations studies. An index to measure perception of this was constructed from three items which ask about the existence of influential cliques of teachers and parents. The questions are:

1. In some schools, when the teachers have anything to say about the running of the school, only the opinions of an inner circle count for much. In other schools, most teachers have a fairly equal voice. Which would you say your school is like in comparison with most schools?
   2.... Dominated by an inner circle of teachers
   1.... Most teachers have an equal voice.
2. Does a teacher or group of teachers who has (have) been in the school a long time have an especially strong voice in influencing the way the principal deals with teachers, students, curriculum or other important school matters?
   0.... No
   1.... Yes
3. Do socially prominent people who have children in the school have an especially strong voice in influencing the way the principal deals with teachers, students, curriculum, or other important school matters?
   0.... No
   1.... Yes

TABLE 5

PERCENTAGE OF TEACHERS PERCEIVING HIGHEST COHESIVENESS WITH OTHER TEACHERS

	Democratic Schools		Autocratic Schools		
Normative Compliance	Percent	N^a	Percent	N^a	Total
Positive	72.5	247	66.9	127	374
Negative	50.9	53	54.2	118	171

N^a = the number of teachers answering the questions in each type of school – the bases from which the percentages were computed.

TABLE 6

PERCENTAGE OF TEACHERS CLAIMING NO CLIQUES EXIST IN THEIR SCHOOL

	Democratic Schools		Autocratic Schools		
Normative Compliance	Percent	N^a	Percent	N^a	Total
Positive	78.9	246	66.1	124	370
Negative	47.1	51	37.8	111	162

N^a = the number of teachers answering the questions in each type of school – the bases from which the percentages were computed.

Index scores can run from 1 to 4 in the direction of most to least favorable. Absence of cliques tends to have a greater relationship to the principal's compliance with the ends norms than to his "democratic" behavior. (See Table 6.)

The subordinate's subjective feelings of freedom to consult the supervisor should relate very closely to his "democratic" behavior, according to human theory, since if a supervisor is "democratic," he attempts to establish effective two way communication. It will be recalled that the Likert scale, which is used to categorize schools as to administrative atmosphere, asks about the behavior of the principal: how often he asks the teacher for advice, how often he associates informally with the teacher, etc. In order to establish the teacher's subjective feelings about consulting the principal (and this should be a result of democratic behavior, not a part of its definition) the teacher was asked:

1. When you have problems, do you feel free to consult with your principal about them?
   1.... Feel very free to consult with principal
   2.... Feel a little hesitant to consult with principal
   3.... Feel very hesitant to consult with principal

The percentage of teachers feeling "very free to consult with principal" is more closely related to the compliance of the principal with the ends norms than to his "democratic" behavior. (See Table 7.)

SUMMARY AND DISCUSSION

It is hypothesized that the compliance of the elementary school principal with what have been called the "ends norms" of the teacher has a stronger relation to various indices of teacher satisfaction than does his "democratic" behavior. There is a tendency for the hypothesis to be upheld in all but one of the six relationships that were examined in this study. The positive compliance of the principal with the "ends norms" ("backing" the teacher and disciplinarianship) has a greater influence than does his "democratic" behavior on the teacher's high evaluation of the principal's leadership, her work satisfaction, her high evaluation of school "clients," her perception of the absence of cliques, and her feelings of freedom to consult the principal. The same trend seems to exist to some degree in the case of feelings of cohesiveness with other teachers in the school. However, teachers in "autocratic" schools who perceive their principals to be non-compliant with the ends norms feel a slightly greater degree

TABLE 7

PERCENTAGE OF TEACHERS FEELING VERY FREE TO CONSULT PRINCIPAL

	Democratic Schools		Autocratic Schools		
Normative Compliance	Percent	N^a	Percent	N^a	Total
Positive	96.8	251	78.0	127	378
Negative	71.7	53	42.7	117	170

N^a = the number of teachers answering the question in each type of school – the bases from which the percentages were computed.

of cohesiveness with other teachers than do teachers in "democratic" schools who perceive their principals to be non-compliant. The findings of this study give no explanation for this tendency, but a guess may be hazarded that the teachers in the "autocratic" schools where the principals are non-compliant are so miserable that they band together for self protection and comfort. The teachers in the "democratic" schools may have at least some two-way communication outlets for communicating their discontent to the principal and perhaps even some hope for having something done about it.

The question arises as to why the "ends norms" of "backing" and disciplinarianship are more important to the teachers than the norm which requires the principal to be "democratic" (hypothesized to be a "means" norm). Every organization may have its own ends norms, and their existence and nature should be discovered in further studies; but why are these particular norms so important to the teacher? The findings of previous research suggest some answers to this question.[16] For one thing, the teacher is a professional or at least (perhaps more accurately) a semi-professional. She is trained to use normative control over her students. However, her students are very young and often are not highly motivated. They may have to be coerced into following the organizational rules "for their own good." The more time the teacher must spend in coercion the less time she has to teach. If the principal cannot be counted on for help with disciplinary problems – if, for example, he frequently sides with the student, the teacher's classroom authority is weakened. When this happens, the students are quick to realize it, and the disciplinary problems become more widespread, more frequent, and more difficult to control. The teacher who is caught in this dilemma cannot teach, which is what she is trained to do. She becomes at best a baby sitter and at worst a jailer. She cannot trust her principal for help, because he may side with the next unruly child or its parents, thus increasing her problems. She then becomes dissatisfied with her job. On the other hand, if the teacher can maintain her authority in the classroom, she can get her job done. The behavior of the principal with regard to the ends norms to a great extent determines whether the teacher's classroom authority will be upheld. The norms which demand that a principal "back" a teacher in front of parents and pupils and that he be a good disciplinarian, are norms which protect the teacher from those with whom she is in constant contact – the children. She is very visible to them and there is a constant threat to her authority unless the principal complies with the "ends norms."

The norms which demand that the principal exhibit democratic behavior protect the teacher's position in the internal organizational structure without immediate regard to the clients (parents and pupils). These norms protect the teacher from the arbitrary use of power by the principal and insure her rights to make her views known. However, because the teacher is not constantly visible to the principal as she is to her students, he is not the threat to her authority that they are, unless his behavior in some way negatively influences the children's concept of her. If the principal complies with the teacher's views regarding the ends norms, she has very little need to communicate with him at all and thus little need for "democratic" behavior on his part. (Although, if the principal's behavior complies with the ends norms, the teacher may like and trust him so much that she will communicate with him freely.) "Democratic" behavior, though itself a norm which teachers hold for their principals, does not appear to be as important to their job satisfaction or to their evaluation of his leadership ability as does his compliance with the "ends norms" of supporting them in front of pupils and parents and of being a good disciplinarian. It does appear to act as a means by which the teachers are allowed to communicate their need for compliance with what they seem to consider more important norms.

*This is an extensive revision of a paper presented at the 1965 Midwest Sociological Society Meetings. The study is part of a larger research project conducted by the Institute for Research in Social Science at the University of North Carolina in Chapel Hill. The director of the larger study is Professor Richard L. Simpson, to whom the author is very grateful for much helpful advice. The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare.

FOOTNOTES

1. Ronald Lippitt and Ralph K. White, "An Experimental Study of Leadership and Group Life," *Readings in Social Psychology*, Editors, Guy E. Swanson, Theodore M. Newcomb, and Eugene L. Hartley (New York: Henry Holt and Company, 1952), pp. 340-355. For some other discussions of leadership effects see: Peter M. Blau, *Bureaucracy in Modern Society* (New York: Random House, 1956); Peter M. Blau and W. Richard Scott, *Formal Organizations* (San Francisco: Chandler Publishing Company, 1962). Blau and Scott have an interesting four-fold classification of formal organizations. The organizations are classified in terms of their prime beneficiary. According to this scheme, the only type of formal organization which necessitates a democratic internal structure is what they call "mutual-benefit associations," where the prime beneficiary is the membership of the organization (p. 43). See also Cecil A. Gibb, "Leadership," *Handbook of Social Psychology*, Volume II: *Special Fields and Applications*, Editor, Gardner Lindzey (Cambridge, Mass.: Addison-Wesley Publishing Company, 1954), pp. 877-917; Kjell Harnvist, *Adjustment, Leadership and Group Relations in a Military Training Situation* (Stockholm: Almqvist and Wiksell, 1956), pp. 120-122; Louis Kriesberg, "Industrial Sociology," *Sociology in the United States of America*, Editor, Hans Zetterberg (New York: UNESCO, 1956), pp. 71-77; Melvin Seeman, *Social Status and Leadership: The Case of the School Executive*, Research Monograph No. 35 (Columbus: Bureau of Educational Research and Service, Ohio State University, 1960); Hanan C. Selvin, *The Effects of Leadership* (Glencoe, Ill.: The Free Press, 1960).
2. For a discussion of assumptions underlying the use of the human relations model see: Robert Dubin, "Psyche, Sensitivity, and Social Structure," *Leadership and Organization*, Robert Tannenbaum, Irving R. Weschler, and Fred Massarik, Editors (New York: McGraw-Hill Book Company, Inc., 1961), p. 403.
3. Amitai Etzioni, *A Comparative Analysis of Complex Organizations: On Power, Involvement and Their Correlates* (New York: Free Press of Glencoe, Inc., 1961), all definitions on page five.
4. Most human relations studies which have focused on the school, for example, have been concerned with the principal-student relationship or the teacher-student relationship, both of which, as Etzioni suggests (*op. cit.*), have elements of coercive power attached to them.
5. A study of a national stratified random sample of 7,150 beginning teachers bears out the above comments. Only 16 percent of the women teachers as opposed to 29 percent of the men expected to teach until retirement. Most of the women planned to become homemakers rather than to remain teachers. For a more complete report of the study see: Ward S. Mason, Robert J. Dressel, and Robert K. Bain, "Sex Role and the Career Orientation of Beginning Teachers," *Harvard Educational Review*, 29 (Fall, 1959), p. 373. Twelve hundred out of a total of 1,250 white elementary school teachers in the present study answered a question about the way they felt regarding their future as teachers. Forty-two and three-tenths percent of them wanted to go into some other kind of work, including becoming full-time housewives.
6. The Bachelor of Arts or Science degrees are held by 73.5 percent of the teachers in the sample. Most of the others (214 or 17.5 percent of the total) have done some graduate work.
7. Refer to sources in footnote 1. Although human relations studies relate "democratic" leadership to worker effectiveness, this paper does not deal with this aspect. It is very difficult to construct a valid measure of teacher effectiveness, especially on the elementary and junior high school levels.
8. Robert K. Merton, *Social Theory and Social Structure* (New York: Free Press, 1957), pp. 339-340.
9. See Peter M. Blau and W. Richard Scott, *op. cit.*, p. 143 for a discussion of the processes by which lenient supervision may result in an extension of the legitimation of the supervisor's authority.
10. Howard S. Becker, "The Teacher in the Authority System of the Public School," *Complex Organizations: A Sociological Reader*, Editor, Amitai Etzioni (New York: Holt, Rinehart, and Winston, Inc., 1961), pp. 243-251. Robert Nelson Bush, *The Teacher-Pupil Relationship* (New York: Prentice-Hall, 1954). Donald C. Moyer, "Teacher's Attitudes Toward Leadership as They Relate to Teacher Satisfaction," (Unpublished Master's Thesis, University of Chicago, 1951). Miriam Wagenschein, "Reality Shock," (Unpublished Master's Thesis, University of Chicago, 1951).
11. In the spring of 1962, when the questionnaires were mailed, North Carolina schools had started to become racially integrated. However, in most cases, the schools were still distinguishable by racial categories. This paper deals only with teachers from white schools since race introduces many new complications into the already complex relationship between administrative atmosphere and aspects of job satisfaction.
12. Twenty-five and eight-tenths percent of the teachers (314) are 29 or younger; 16.2 percent are between 30 and 39 (197); 23.0 percent are between 40 and 49 (280); 27.4 percent (333) between 50 and 59; and 7.6 percent (92) are 60 years old or older. The total number of respondents to this question is 1216.
13. All but item 6 in this scale are selectively drawn from Melvin Seeman's interesting monograph, *Social Status and Leadership: The Case of the School Executive*, Research Monograph No. 35 (Columbus: Bureau of Educational Research and Service, Ohio State University, 1960). The question may be raised as to why the teacher's perceptions of the principal's behavior were obtained, rather than the principal's self-reports of their actual behavior. An assumption is being made here that the perceptions an individual has about a situation are more important in shaping his attitudes concerning the situation

than are the "objective facts." Further, since all of the respondents from each school rate the principal on the same items, the average of the collective responses should serve as a basis of inference of actual behavior. The self-report of the principal may be honest in that he is reporting what he thinks he does, but his actions may not be perceived or reacted to in the way in which he intended.

14. A score of 3 means perception of the highest principal compliance on all three items; a score of 4 means perception of highest principal compliance on two items and above average compliance on one; a score of 5 means either perception of highest compliance on one item, and above average on two, or average on one and outstanding on two. The cutting point is drawn at a score of 5, because it is believed that a potential "average" rating on more than one item would be too low a rating for the positive category. Therefore, a score of 5 or under is placed in the positive category and a score of 6 or over is placed in the negative category. Because the literature stresses the teachers' wishes for principals who "back them up" and are good disciplinarians, and because the literature does not give special importance to any one of these behavioral aspects, the items are given equal weight.

15. This wording was a result of a pretest in which most complaining teachers said such things as "I don't like their attitude."

16. See especially, Becker, op. cit.; Bush, op. cit.; Etzioni, A Comparative Analysis, op cit.; and Wagenschein, op. cit.; and Moyer, op. cit.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

THE BORDERLINE STUDENT AS A MEMBER OF A SEPARATE SCHOOL GROUP

DONALD R. HETZNER

Medina School System
New York

ABSTRACT

Although the Medina, New York, Prevocational Program for "slow learners" had been in operation for 11 years there was no adequate proof that the Prevocational section varied significantly, on the basis of mean intelligence, from either the Special Education or academic section of the school system. Intelligence quotient scores were gathered and mean intelligence, standard deviation, and confidence interval were established for each of the sections. Null hypotheses were then used to test for significant differences between the central tendencies. This established that Prevocational, Special Education, and the academic section were discrete groups.

Educational facilities drawing their enrollment from a representative section of the general population will, under Medina criteria, have approximately 17-20 percent "slow learners" in their school population. If the needs of these borderline students are to be fulfilled it is the responsibility of the schools to design curricula tailored to a group falling midway in a progression between Special Education and academic classes.

SINCE 1955 the Medina, New York, School System has offered a program for teenagers whose ability to achieve the standards of academic courses is doubtful. For 7 years the program was offered only in the seventh and eighth grades; in 1961 it was expanded to include the ninth grade of the Senior High School. After 1962 the Prevocational Program was expanded a year at a time until, by 1965, it covered grades seven through twelve and graduated its first class. The original program, and its expansion to include junior and senior high school, was based on the idea that the maintenance of a comprehensive community high school dictated that the needs of all significant groups of students be met. Thus, those students who are, in a general sense, "academic slow learners" must have a curriculum tailored to their needs--a course of study that holds them in school until they are adequately prepared to find and keep a job suited to their abilities (5).

MENTAL ABILITY AND READING DEFICIENCY

The Medina Prevocational Program is a terminal education plan admitting only students who, for want of a better term, are considered "slow learners." These are students selected because their success in an academic program is impeded by reason of limited mental ability (intelligence quotients in the range 75-90), restricted background, and/or a serious reading deficiency. These students, 72 percent of whom have intelligence quotients in the range 75-90, are considered "slow" only in regard to academic capacity; in other areas of endeavor they may rate average or better. Of the three factors mentioned mental ability and reading level, as measured by standardized tests, are considered the most unbiased estimators used to evaluate and place students, particularly as all past intelligence and reading tests are taken into account when considering a student for the program.[1]

At present, of the 135 students in grades 7-12 with intelligence quotients in the range 75-90, ninety-eight are in the Prevocational Program. Of the 168 students participating in the plan only one student was reading at grade level when admitted. From all indications of a study now in progress there is a positive though non-linear relationship between below-normal intelligence and reading deficiency.[2] The remaining 28 percent of the students in the program have a mean intelligence quotient of 95.83, above the cut-off point of 90 but placed in the program because of what are considered irreversible reading deficiencies. When dealing with students having intelligence quotients in the range 91-100 the Prevocational Department has found that pupils reading more than two years below grade level are so severely hindered that any success in an academic program is very doubtful. There are several empirically verifiable conclusions that can be drawn from the existing data:

TABLE 1
ACHIEVEMENT TESTS (Median Number of Months Below Grade Norm)

Grade	Reading Iowa Basic Skills	Reading Stanford	Arithmetic Iowa Basic Skills	Composite Iowa Basic Skills	Age Difference In Months From Median Age of Rest of Grade
7	-19	-16	-8	-14	+12
8	-14	-19	-10	-14	+14
9	-12	-18	-6	-10	+13
10	-14	-18	-10	-12	+11
11	-14	-17	-10	-12	+11
12	-15	-17	-8	-12	+15

From "The Prevocational Program in Grades 7-12 of Medina Central School, Sixth Progress Report" (Medina, New York: Medina Central School, June, 1967) (Mimeographed).

a student of normal intelligence can function in an academically oriented class even with a severe reading deficiency, but it is impossible for a student with limited intelligence and a reading disability to compete in an academic class. Reading ability is extremely critical for students who are on the borderline of entering the program by the results of their scores on intelligence tests. These are pupils who have intelligence quotients in the range, roughly, of 91-100 and severe reading problems. There is some question within the program regarding what is to be considered a reading problem serious enough to impair any chance of success in an academic program. The usual practice is to consider a student with an intelligence quotient of over 90 for the program only if he is a year or more behind in reading, has failed a grade and perhaps is in danger of failing again. By following these entrance requirements closely students are placed in the program because of limited mental ability and an irremedial lack of reading skills, not because they are "underachievers"(4). Table 1 presents an overall view of the achievement levels and age differences between the Prevocational classes and the grade norms.

THE NEED TO VERIFY THE REASON FOR THE PLAN'S EXISTENCE

Though the Medina program for "slow learners" has existed for 11 years, there has never been an adequate statistical proof that the Prevocational and Academic sections are significantly different. This also seems to be the case in other schools where academically slow students are grouped separately. It does not suffice as a proof merely to state that students in a slow group have intelligence quotients between 75-90 inclusive, or that the mean intelligence of the group is say, 87. In the first case the composition of groups will vary from year to year and/or group to group, and the intelligence curve or the sample may be skewed toward the lower or, more likely, the upper boundary of the range. A mean intelligence quotient in isolation, as in the second case, is valueless unless used for prediction or in valid comparison with other groups. Thus it is necessary to establish a confidence interval for the mean intelligence of the "slow learner" group and a statistical measure, in this case mean intelligence of all Prevocational students in Medina, which can be used to test hypotheses about the "slow learner" group.

THE ERROR OF OBSERVATION

The original impetus behind the Prevocational Program came from teachers who recognized the inability of certain students to cope with an academic program--still--they were pupils who definitely did not belong in Special Class. Because academic courses were often geared toward college entrance, it was felt that such study did not meet the needs of students whose last aquaintance with formal education would be, at most, high school. The solution was to place these students in a program suited to their abilities, a program where the formation of socially accepted attitudes, habits and the learning of occupational skills took precedence over abstract academics. The factual knowledge and skills required of the students were modified accordingly.

A cursory inspection of a Prevocational group would seem to indicate nothing more than a group of normal teen-agers. In fact, one would see a class that, because of the stress placed on acceptable habits and attitudes, is less given to fadish dress and speech modes than other groups. However, because of the seeming normalcy of the class many observers not aquainted with the situation would doubt the fact that this is a group different from any number of poor academic classes. The truth is though that, as a group, the Prevocational students' abilities to pursue abstract reasoning, comprehend analogies, and make inferences from factual data falls approximately midway between that of students in a regular class and Special Education. All the above factors, combined with a nearly one hundred percent incidence of below grade level reading, produces to view an outwardly average looking group of teenagers. But, concealed under overt appearances are individuals who, unless guided toward social responsibility, become tomorrow's unemployables. Extrapolating from the data at hand--these are individuals who, in most school systems, are not recognized as being members of a valid group and do not have their life needs fulfilled. In the following sections null hypotheses will be used to test for significant differences in mean mental ability between the Special Education, Prevocational, and Regent's sections of the Medina school system. "Slow learners" form a group that, though overlooked, still exists and whose needs cannot be met by having its members scattered in academic groups and subject merely to individual counseling. They are not isolated cases--they are members of a valid group that exists in every school that draws its students from a normally distributed population and should be treated accordingly (2).

PURPOSE OF THE STUDY

Students in the Prevocational Plan are selected and placed by criteria evolved through long term observation and research. Selection procedure, however, is not the point of this paper. Rather, the center of interest is the question: Are Prevocational students, once selected and placed in a homogeneous group by a rational and coherent set of standards, the nucleus of a group significantly different in mental ability from the Special Education and academic sections of the school system (4). It is realized that the make-up of a group is the function of the selection criteria. In this case, however, there is a group already in existence. Therefore, the fallacy of positing a group, setting the selection criteria, and then testing the group is avoided. Although the composition of a group will vary directly in relation to the selection procedure, if the selection process is held constant, statistical technique can be applied to determine whether a Prevocational class varies significantly from other--more standard--classes.[3]

A NOTE ON METHODOLOGY

Data for this study and for a larger monograph of which it is a part, were taken from the guidance files of the Medina Junior and Senior High Schools. The intelligence quotient scores for the students in Regent's and Prevocational sections are the results from Otis Beta tests and those for Special Education are based on Stanford-Binet and WISC tests. In all cases the most recent test score was used. After the data were gathered and catagorized, coding techniques were used to determine mean intelligence and standard deviation for each of the four sections in the school system--Special Education, Prevocational, Regent's, and Advanced. The Advanced group was treated separately from the Regent's section for the reason that they follow a course of study with requirements beyond the state-mandated minimum standards. Within the System they are treated as a separate section and are treated so for the purposes of this study.[4] Moreover, by treating the Advanced students as a group separate from the Regent's section, a truer picture of the student following the state-mandated requirements is arrived at. And, this student following the regular academic pattern of school life is the person to whom we wish to direct part of our attention, comparing him with the student who follows the Prevocational curriculum and, in turn, comparing the Prevocational pupil to the student in Special Education. Is the central tendency of the Prevocational group in Medina, as measured by mean intelligence, statistically different from those of Special Education and Regent's ? The lack of a statistically significant difference between Prevocational and either of the other two sections would leave the value of the Prevocational Program open to question.

THE MODEL

In an idealized situation, with intelligence quotient as the sole criterion for grouping, the terms of the problem are: intelligence quotient cut-off points of 75 and 90 imply that students with intelligence quotients below 91 are classed as "slow learners"; those above 90 are deemed able to handle the abstract material of Regent's classes; and those below 75 are "educable" in a Special Education class. "Slow learners" enter the Prevocational Program, "normal learners" are placed in Regent's, and "educables" enter Special Education. Therefore there are no "slow learners" in the academic section, no "normal learners" in the Prevocational Program, and no "educables" in either of the other sections--the three groups are dissimilar on the basis of different mean intelligence scores within the groups. The varying situations and possibilities may be outlined as follows:

Special Education and Prevocational

1. Students with intelligence quotients below 75 are "educables" and those with intelligence quotients above 74 and below 91 are "slow learners."
2. "Educable" pupils enter Special Education and "slow learners" enter the Prevocational Program.
3. There are no "educables" in Prevocational and no "slow learners" in Special Education.
4. Special Education and Prevocational Program students form two distinct groups.

Prevocational and Regent's

1. Students with intelligence quotients below 91 are "slow learners" and those above 90 are "normal learners."
2. "Slow learners" take Prevocational courses and "normal learners" enter the Regent's program.
3. There are no "slow learners" in Regent's and no "normal learners" in Prevocational.
4. Regent's and Prevocational Program students form two distinct groups.

The above syllogisms, restates as testable hypotheses, are in the form:

H_1 - Mean of Prevocational equals mean of Special Education

H_2 - Mean of Prevocational does not equal mean of Special Education

H_3 - Mean of Prevocational equals mean of Regent's

H_4 - Mean of Prevocational does not equal mean of Regent's

CONFIDENCE INTERVALS

The data of Table 2 were used to establish confidence intervals for the mean intelligences of Special Education, Prevocational, and Regent's and to test the foregoing hypotheses.

TABLE 2
SUMMARY DATA

	Mean Intelligence	Standard Deviation
Special Education	65.58	10.26
Prevocational	87.55	10.12
Regent's	109.04	23.54
Advanced	122.04	16.30

Prevocational

$85.90 < \mu < 89.20$ is established as the 95 percent confidence interval for the mean intelligence of the group formed by all Prevocational students in grades 7-12. In other terms, using this research methodology, the probability is .95 that the interval 85.90 - 89.20 will contain μ, the population mean, as estimated by $\bar{x}$, a sample mean used as the point estimate of μ.

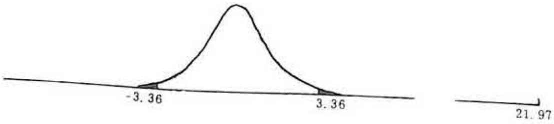
Regent's

$107.31 < \mu < 110.76$ is the 95 percent confidence interval for the mean intelligence of the Medina Regent's group in grades 7-12.

Special Education

The establishment of the interval $62.70 < \mu < 68.50$ as containing the intelligence mean of all Special Education students leads to the next and most important section of the research--a proof that, on basis of mean intelligence, Special Education, Prevocational, and Regent's are discrete groups.

H_2 IS PROVED: MEAN INTELLIGENCE OF PREVOCATIONAL DOES NOT EQUAL MEAN INTELLIGENCE OF SPECIAL EDUCATION



As a first step in proving that the mean intelligence of Prevocational differs from that of Special Education the sample standard deviation $\sigma \bar{x}_1 - \bar{x}_2$ will be computed by combining the standard deviations of Prevocational and Special Education to arrive at a sample estimator.

$\bar{x}_1$ = sample mean of Prevocational

$\bar{x}_2$ = sample mean of Special Education

$\sigma \bar{x}_1 - \bar{x}_2$ = 1.68 Standard Deviation

Now the hypotheses $H_1 : \bar{x}_1 = \bar{x}_2$; $H_2 : x_1 \neq x_2$ will be tested. The critical region for the test is based on equal tail areas under a normal curve with $\alpha = .05$. Thus, the critical region is that part of the horizontal axis more than $2\sigma \bar{x}_1 - \bar{x}_2 = 3.36$ units from 0 (zero) where 0 represents the mean intelligence of Prevocational -- 87.55. As the two samples of 142 and 50 students yielded the value $\bar{x}_1 - \bar{x}_2 = 87.55 - 65.58 = 21.97$ and 21.97 falls in the critical region the hypothesis $\bar{x}_1 = \bar{x}_2$ is rejected.

The central tendencies of the Prevocational and Special Education groups in grades 7-12, as measured by mean intelligence, are significantly different.

H_3 IS PROVED: MEAN INTELLIGENCE OF PREVOCATIONAL DOES NOT EQUAL MEAN INTELLIGENCE OF REGENT'S

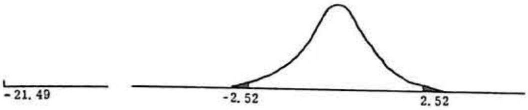
Turning to H_3: $\bar{x}_1 = \bar{x}_3$ and H_4: $\bar{x}_1 \neq \bar{x}_3$ the same procedure employed in the Prevocational-Special Education proof will be used to test these hypotheses; by combining the standard deviations of the mean intelligences of the two groups $\sigma \bar{x}_1 - \bar{x}_3$ will be established.

$\bar{x}_1$ = sample mean of Prevocational

$\bar{x}_3$ = sample mean of Regent's

$\sigma \bar{x}_1 - \bar{x}_3$ = 1.26 Standard Deviation

The critical region is based on equal tail areas under the normal curve for $\bar{x}_1 - \bar{x}_3$ with $\alpha = .05$ -- that part of the horizontal axis more than $2\,\sigma\,\bar{x}_1 - \bar{x}_3 = 2.52$ units from zero where 0 is, again, the mean intelligence of the Prevocational group -- 87.55. With the two samples of 142 and 630 students yielding the value $\bar{x}_1 - \bar{x}_3 = 87.55 - 109.04 = 21.49$ the hypothesis $\bar{x}_1 = \bar{x}_3$ is rejected as -21.49 falls in the critical region.



The central tendencies of the Prevocational and Regent's groups in grades 7-12, as measured by mean intelligence, are significantly different.

EVALUATING THE PROGRAM

Research on the Medina Prevocational Program has been largely aimed at evaluating the effects of the program on the students. Two of the more important studies are Nunzio T. Maiorana "Comparative Analysis of Attitudes," and George C. Barna "The Holding Power of the Prevocational Program."

Nunzio Maiorana, in a long-term study, has concentrated his attention on determining whether the attitudes of Prevocational students change after entering the program (3). In the study he deduces attitudes by observing the choices students make in their overt behavior. Maiorana uses the sixth grade records of Prevocational students as his base, and finds that after entering the program attendance is up 54 percent; the amount of homework turned in approximates the percentage turned in by a random sample of students in academic classes; distribution of grades compares closely with academic classes; and the number of behavior problems, as determined by examining anecdotal records, are down 60 percent.

George Barna, prior to the graduation of the first Prevocational class, compared the dropout rate of students with intelligence quotients of less than 95 who were in regular classes to the dropout rate of pupils in the Prevocational Program (1). He ascertained that the Prevocational Program exerted a holding power 12 percent greater than other programs in the system. That is, a 20 percent dropout rate in Prevocational as compared to a 32 percent dropout rate for a comparable group of students who were in other school programs. Mr. Barna is now in the process of comparing the dropout rate of the Prevocational Program to the statewide rate for students with borderline intelligence quotients. For the purposes of his study an intelligence range of 75-95 is being used. The updated studies of Maiorana and Barna will appear, along with other descriptive statistics, in the "Prevocational Program in Grades 7-12 of Medina Central School, Sixth Progress Report," copies of which can be obtained from the Medina Central School System. (6)

THE APPLICABILITY OF THE STUDY

What is to be gained from an isolated study of this nature? Results can be generalized to encompass a larger population and geographic area. It can be assumed, on the basis of other studies, that "slow learners" are normally distributed and comprise 17-20 percent of the general population so, approximately these same results will hold for samples of normally distributed populations if selection criteria are held constant. Thus, educational facilities which draw their enrollment, in an unbiased manner, from a representative section of the general population, will contain a proportionate number of "slow learners" within the schools involved. If this hypothesis holds true then it becomes the responsibility of school systems to consider designing curricula tailored to the needs of "slow learners" who, on the basis of mean intelligence, form a group midway in a progression between Special Education and academic classes.

FOOTNOTES

1. Results on tests of mental and reading ability are easily quantifiable while assignments based on "limited background" are extremely subjective. In a public school system, where it is impossible to examine a student's home life in detail, an estimate of deprivation must be made. In the majority of cases, however, there is a colinear relationship between two or all of the factors.

2. From results gathered during an experimental program of group remedial reading for the "slow learner" it was found impossible to predict the amount of reading disability or a possible increase in reading ability on the basis of measured intelligence.

3. Selection procedure determines the nature of any group. In the case of the Prevocational Program the assigned cut-off points of 75 and 90 determine that, almost assuredly, the mean intelligence of the group will fall in this range. Whether or not the mean intelligence is significantly different from that of other groups must be determined statistically.

4. An "Advanced" student is a pupil having one or more classes in which the content extends beyond the state-mandated minimum requirements.

REFERENCES

1. Barna, George C. "The Holding Power of the Prevocational Program," (Medina, New York: Medina Central School), (Mimeographed).

2. Cardiello, Conrad I. "Prep Program Provides Placement Opportunities for Underachievers," New York State Education, (November, 1966), pp. 41-43.

3. Maiorana, Nunzio T. "Comparative Analysis of Attitudes," (Medina, New York: Medina Central School), (Mimeographed).

4. "Minutes of the Prevocational Department," (Medina, New York: Medina Central School, February, 1965), (Mimeographed).

5. "The Prevocational Program in Grades 7-12 of Medina Central School, Fifth Progress Report," (Medina, New York: Medina Central School, February, 1965), (Mimeographed).

6. "The Prevocational Program in Grades 7-12 of Medina Central School, Sixth Progress Report," (Medina, New York: Medina Central School, June, 1967), (Mimeographed).

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

GENERATING A PROJECTED CRITERION OF GRADUATE SCHOOL SUCCESS VIA NORMATIVE JUDGMENT ANALYSIS[1]

SAMUEL R. HOUSTON[2]
Institute for Development of Educational Activities

ABSTRACT

This study was designed to investigate the extent to which a policy regarding graduate admission standards exists among selected graduate faculty members at Colorado State College. Twenty judges utilized Normative Judgment Analysis techniques to generate a projected criterion of graduate school success on profile data for thirty randomly selected doctoral graduates. The results of the study indicate that essentially one policy was being expressed by the judges.

A SEARCH of the literature reveals that the number of published studies in predicting graduate school success is small and nearly all sample sizes have been rather limited. In general, the studies have grouped graduate students together and have ignored the variable of the major department in which the graduate students are enrolled. Typically, graduate grade point average (GPA) is utilized as criterion of success. Criticisms of using the GPA as the dependent variable are made both in terms of its appropriateness as well as with the limited range of grades (variability) in graduate school. Nevertheless, graduate schools are faced with the serious problem in the identification of potentially talented graduate students and must rely on profile admission data in the decision-making process.

A potentially useful technique which avoids the criticisms resulting from using the GPA as the dependent variable is Normative Judgment Analysis. Normative Judgment Analysis is a special adaption of a technique developed by Bottenberg and Christal (1) which groups criteria in terms of the homogeneity of their prediction equations. This approach is, in turn, a special application of a still more general hierarchial grouping developed by Ward (3). Normative Judgment Analysis (JAN) is a multiple linear regression approach used to determine the relative importance of the various profile items for each judge in the expression of his policy. In the case of admitting students to graduate school, it allows the admitting school to make a decision whether to accept a given student on the basis of a policy which has been expressed by JAN. The policy is a statement of the weights to be applied to scores obtained on the various profile items for each student seeking admission. It clearly identifies a group position on the relative importance of the various profile variables utilized in making decisions about the admission of potential students.

What is the sequence of steps in JAN? To illustrate the process, consider a situation in which a judge is presented with a number of applicants for a job where each applicant has been given a score on a number of different dimensions (profile items) thought to be relevant to the job. Instead of following the usual procedure of asking each judge to determine the relative importance of each of the profile items so that a proper weighting of the variables can be obtained in forming a selection battery, each judge instead is asked to examine each applicant and give him an overall score that will serve to rank the applicants in an order of merit. The assigned score for each applicant is then utilized as a projected criterion score where the profile items or dimensions are used as predictors. This results in a series of zero-order validity coefficients which relate each individual predictor (profile item) to the criterion and which in turn can be used in obtaining a multiple regression equation for that judge which identifies his policy. Geometrically, the raw score weights obtained in this equation determine the hyperplane which is a statement of policy for that judge and the resulting RS (the magnitude of the obtained squared multiple correlation coefficient) is an indicator of the intrarater consistency of his judgments. With this approach it becomes possible to present a number of judges with the same set of stimuli and consequently obtain a policy equation for each judge.

Once these equations, or profiles, have been obtained from all the judges, a grouping procedure may be applied which identifies the minimum number of policies that are essentially being expressed, as well as the areas of disagreement. This criterion-grouping technique is probably best described by quoting from Christal as follows:

> Beginning with these equations, the criterion-grouping technique is employed to define

areas of agreement and disagreement among the board members. As the first step, a single value of R^2 is computed which indicates the overall predictive efficiency obtained when a separate least-squares-weighted regression equation is used for each rater. During the second step, each equation is compared with every other equation. In this way, the two raters are located who are in closest agreement concerning how the selection variables should be weighted to determine the relative acceptability of applicants. These two raters have the most homogeneous regression equations. The computer then determines the single equation that best represents the joint policy of these two raters. It also indicates the overall loss in predictive efficiency that results when the original N equations are replaced by the best set of N - 1 equations. This is accomplished by comparing the magnitude of the overall R^2 in the two instances.

In subsequent operations, the program systematically reduces the number of prediction equations (or rater clusters) by one at each step so as to minimize the loss in overall predictive efficiency. In each instance, the computer prints out an overall R^2, as well as the single equation for each cluster which best represents the joint policy of all raters in that cluster. At the last step, a single prediction equation is generated so as to express the joint policy of all raters with as little error as possible.

Examination of the loss in overall predictive efficiency at various stages of the grouping process will help to identify the number of rating policies existing among board members. For example, if very little overall predictive efficiency is lost in going from ten to two equations, but a considerable amount of efficiency is lost in going from two to one equation, there is strong evidence that two different rating policies are operating. That is to say, some of the board members are using one weighting system and the rest a different weighting system. Of course the hope is that a single equation can be found which predicts the criterion values assigned by all raters. In this case the board could be dismissed without further ado, since their single agreed-upon policy would have been determined. However, should more than one rating policy be found to exist, the board is not asked to arbitrate the difference in weights. Rather, the several equations are used to rank the applicant sample, and the board is asked to arbitrate the differences in the rank position of cases resulting from application of these several equations. Once a single agreed-upon rank-ordering of the sample has been obtained, it is a simple matter to determine the appropriate set of weights for the final selection composite.

The JAN technique enables the investigator to make a complete analysis of interrater agreement. As a matter of fact, results of the first step alone indicate the amount of agreement between each rater and every other rater (2: 4).

The JAN procedure seems to have numerous advantages recommending its use. The most important is that it enables the investigator to capture the policy of a judge in a very clear and methodical manner. In addition, it is "objective" in the sense that each judge is asked to render a decision based on available data rather than on individuals or personalities. Further, JAN permits the use of nonlinear terms in the regression model which means it is essentially unrestricted mathematically. Finally, another obvious advantage is that its basic judgment is probably a much more realistic assignment for judges than the typical situation which requires the judge to ascribe the various weights to each predictor variable.

DESCRIPTIONS OF THE VARIABLES (Table 1)

1. Normative-JAN criterion

The rating of each of the Ss was pentamerous coded, 1 identifying those students whose projected graduate success placed them into the bottom 20th percentile of successful graduate students at Colorado State College; 2 identifying those students whose projected graduate success placed them into a percentile between the 20th and the 40th; 3 identifying those students whose projected graduate success placed them into a percentile between the 40th and 60th; 4 identifying those students whose projected graduate success placed them into a percentile between the 60th and the 80th; 5 identifying those students whose projected graduate success placed them at or above the 80th percentile.

2. Colorado State College master's degree

Whether or not a master's degree was received from Colorado State College was binary coded for each of the Ss, 1 identifying those students who obtained their master's degree from Colorado State College; O identifying those students who received their master's degree from some school other than Colorado State College.

3. Cumulative grade point average

Cumulative grade point average was recorded on a five-point scale, 5 points identifying an A; 4 points identifying a B; 3 points identifying a C grade; 2 points identifying a D grade; 1 point identifying an F grade. It included all bachelor's and master's degree work taken previous to undertaking doctoral degree studies.

4. Years between B. A. and M. A. degree

5. Rating on Interview

The rating on the interview of the Ss was recorded on a four-point scale; 1 identifying those students most unlikely to succeed; 2 identifying those students who are doubtful; 3 identifying those students who are likely to succeed; 4 identifying those students who are most likely to succeed. This rating represents the concensus of the graduate committee appointed to interview each applicant for the doctoral degree program.

6. Recommendation on Interview

Coded 1 if the S was recommended to continue in the doctoral degree program, other wise O.

TABLE 1

LIST OF VARIABLES

Number	Variable	Abbreviation
1	Normative-JAN criterion	NJ
2	Colorado State College master's degree	CSCM
3	Cumulative grade point average	CuGPA
4	Years between B. A. and M. A. degree	YrBD
5	Rating on interview	RaI
6	Recommendation on interview	ReI
7	Sex	Sex
8	GRE-Verbal	GRE-V
9	GRE-Quantitative	GRE-Q
10	GRE-Social Science	GRE-SS
11	GRE-Humanities	GRE-H
12	GRE-Natural Science	GRE-NS
13	GRE-Advanced Education	GRE-Ed
14	English Usage Test score	EnU
15	Education major	EdM
16	Science major	SM
17	Psychology major	PM
18	Business major	BM
19	Music major	MM
20	Industrial Arts major	InM
21	Physical education major	PEM
22	Other major field	OtM

7. Sex

Coded 1 is the S was a male, O for a female.

8. Graduate Record Examination - Verbal score

9. Graduate Record Examination - Quantative score

10. Graduate Record Examination - Social Science score

11. Graduate Record Examination - Humanities score

12. Graduate Record Examination - Natural Science score

13. Graduate Record Examination - Advanced Education score

14. English Usage Test score

The English Usage Test is a special test developed and standardized at Colorado State College designed to measure written communication skills. Results were reported as percentile scores.

15. Education Major

Coded 1 if the S was an education major, otherwise O.

16. Science Major

Coded 1 if the S was a science major, otherwise O.

17. Psychology Major

Coded 1 if the S was a psychology major, otherwise O.

18. Business Major

Coded 1 if the S was a business major, otherwise O.

19. Music Major

Coded 1 if the S was a music major, otherwise O.

20. Industrial Arts Major

Coded 1 if the S was an industrial arts major, otherwise O.

21. Physical Education Major

Coded 1 if the S was a physical education major, otherwise O.

22. Other Major Field

Coded 1 if the S was a social science, English or general plan major, otherwise O.

PROCEDURES

Twenty judges, representative graduate faculty members from Colorado State College, were requested to respond to the profile data on 30 randomly selected doctoral graduates from Colorado State College (1963-1966). Each judge was asked to place each student on the basis of the profile information into one of five categories. A score of five indicated that the judge thought the student would probably graduate in the top fifth of Colorado State College graduates. Similarly, a score of one indicated a judgment that the student would graduate in the lowest fifth of Colorado State College graduates. The purpose was to ascertain indirectly a group policy on the part of the faculty in terms of the predictor variables. Sixteen or 80 percent of the judges responded.

RESULTS AND DISCUSSION

A multiple linear regression equation was developed utilizing the data obtained from the 16 judges with 21 predictor variables to predict a criterion variable of ranked graduation category. The present analysis was restricted to an investigation of two questions: a) the efficiency of the 21-variable system for predicting the ranked graduation category and b) the efficiency of the prediction when certain groups of predictors were removed from the system.

Essentially, the three groups of predictor variables investigated were : a) Biographical items (variables 2-7 inclusive); b) Test data (variables 8-14 inclusive); and c) Major subject variables (variables 15-22 inclusive).

TABLE 2

INTERCORRELATIONS

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
NJ 1																						
CSCM 2	-37																					
CuCPA 3	-10	31																				
YrBD 4	-01	-21	-08																			
RaI 5	05	07	09	30																		
ReI 6	01	19	03	-16	06																	
Sex 7	04	-33	-10	12	-04	-22																
GRE-V 8	38	17	36	-08	27	09	-19															
GRE-Q 9	27	-32	31	-53	-25	35	-12	36														
GRE-SS 10	53	11	27	-16	09	12	-09	76	46													
GRE-H 11	16	03	27	-30	04	21	-45	56	22	42												
GRE-NS 12	49	20	29	-21	13	24	-06	75	61	82	49											
GRE-Ed 13	16	28	29	-10	31	31	-25	67	25	64	30	54										
EnU 14	37	00	04	12	18	20	-24	58	17	47	46	49	26									
EdM 15	09	-05	27	13	22	21	-10	14	12	33	11	14	40	02								
SM 16	11	17	12	15	28	13	05	20	05	14	12	34	04	10	13							
PM 17	09	21	-09	-63	-17	30	-11	25	31	18	27	27	30	01	30	-12						
BM 18	-59	31	19	17	18	-17	-14	-24	-26	-39	-11	-33	-21	-05	-17	-05	-16					
MM 19	-66	18	35	08	-26	00	-06	-22	-22	-39	18	-34	-28	-19	00	-05	-17	37				
InM 20	-31	-03	-05	01	04	-01	01	-07	-03	-27	-10	-14	-22	00	-01	-05	-18	14	18			
PEM 21	-71	45	09	10	-05	00	-06	-23	-12	-36	-25	-34	-01	-11	00	-17	-17	38	47	18		
OtM 22	-83	-95	38	-06	01	22	-17	-07	-03	-28	11	-20	10	-13	22	-03	-03	58	72	33	72	

The intercorrelation coefficients (Table 2) among the 22 variables are product-moment coefficients. Because some of the variables are binary coded, some of the coefficients are phi's and some are point-biserial's.

The predictor variables were selected because they represented readily accessible types of information which, it was believed, might have some predictive relationship to the policy expressed by the judges. The coefficients in column 1 of Table 2 show these relationships. Reference to Table 2 shows that the three highest zero-order correlations between predictor variables and the criterion are Other Major Field (-.83), Physical Education Major (-.71), and Music Major (-.66).

The investigator used multiple linear regression, according to Ward (4: 204-237), to determine the unique contribution of proper subsets of the predictor variables, (2-22), to the prediction of the criterion, Normative-JAN rating. The contribution of a set of variables to prediction may be measured by the difference between two squares of multiple correlation coefficients (RSs), one obtained for a regression model in which all predictors are used, called the full model (FM), and the other obtained for a regression equation in which the proper subset of variables under consideration has been deleted; this model is called the restricted model, (RM). The RS for the RM can never be larger than the RS for the FM. The difference between the two RSs may be tested for statistical significance with the variance ratio test. The hypothesis tested states, in effect, that these variables contribute nothing to the determination of the expected criterion values that is not already available in the restricted prediction system.

There are several possible interpretations of the unique contribution of a variable to the prediction of a criterion. If a variable is making a unique contribution, then knowledge of that variable furnishes information about the criterion. Also, if a variable is making a unique contribution, then two Ss, who are unlike on the variable but who are exactly alike or are matched on the other predictors, will differ on the criterion.

The results of the regression analysis are summarized in Tables 3 and 4. Table 4 indicates that

TABLE 3

MEANS AND STANDARD DEVIATIONS (N = 30)

Variable		Mean	Standard Deviation
1	NJ	2.80	2.15
2	CSCM	.48	.58
3	CuGPA	3.96	.46
4	YrBD	7.99	3.59
5	RaI	1.87	.76
6	ReI	.67	.48
7	Sex	.93	.24
8	GRE-V	497.31	92.78
9	GRE-Q	498.86	115.96
10	GRE-SS	523.75	85.00
11	GRE-H	511.96	76.13
12	GRE-NS	514.20	100.52
13	GRE-Ed	551.06	78.81
14	EnU	65.32	25.83
15	EdM	.33	.47
16	SM	.03	.18
17	PM	.30	.46
18	BM	.08	.31
19	MM	.11	.43
20	InM	.07	.26
21	PEM	.11	.43
22	OtM	.01	.27

TABLE 4

PROPORTIONS OF VARIANCE ATTRIBUTABLE TO GROUPS OF VARIABLES BELIEVED TO BE ASSOCIATED WITH NORMATIVE JAN CRITERION

PREDICTOR-Variable Group	Total Contribution Proportion (RS)	Unique Contribution Proportion	Multiple Correlation with Criterion
Model 1 (2-22)	.8580		.93[b]
Model 2 (2-14)	.5928		.76[b]
Model 3 (2-7, 15-22)	.8127		.90[b]
Model 4 (8-22)	.8203		.91[b]
Major Subject Field (15-22)		.2652[a]	
Test Data (8-14)		.0453[a]	
Biographical Data (2-7)		.0377[a]	

[a] All proportions reported as unique contributions are significant at the .01 level for N = 480. In computing F values, it was assumed that one parameter was associated with each variable in the prediction system. The degrees of freedom for the number of predictors were determined by the number of variables given an opportunity to contribute to the prediction.

[b] Significant at the .01 level.

the 21-variable composite accounts for .8580 of the total criterion variance which is significant at the .01 level. A multiple correlation using the 21 variables as predictors of the Normative-JAN rating was .93. This is not only significant at the .01 level, but it is exceptionally high from a predictive point of view as approximately 86 percent of the variance is accounted for when the 21 variables are used as predictors. The high RS (.8580) indicates that the interrater agreement is exceptionally high and that there is essentially one policy being expressed by the judges using the Normative JAN technique.

In model 2 a 13-variable composite (2-14) was tested for predictability in which the eight variables (15-22) which identified the major subject field of doctoral study were omitted as predictors. About 59 percent (.5928) of the criterion variance is estimated to be attributable to the 13 variables which essentially consist of biographical items and test data. The difference between RS value for the FM system (.8580) and the RS value for the restricted model, FM - (15-22), yields an estimate of .2652 for the unique contribution of variables (15-22) as a group. The loss in predictive accuracy when major subject field data are not considered seems sufficient to be of practical importance. The unique contribution of the eight predictor variables (15-22) involved with identifying the major subject field was significant at the .01 level.

In model 3 a 14-variable composite (2-7, 15-22) was tested for predictability in which seven variables (8-14) which yielded test data information were omitted as predictors. About 81 percent (.8127) of the criterion variance is estimated to be attributable to the 14 variables which essentially consist of biographical items and major subject field data. The difference between RS value for the FM system (.8580) and the RS value for the restricted model, FM - (8-14), yields an estimate of .0453 for the unique contribution of variables (8-14) as a group. The loss in predictive accuracy when the test data (8-14) are not considered seems sufficient to be of practical importance. The unique contribution of the seven predictor variables involving test data information (8-14) was significant at the .01 level.

In model 4 a 15-variable composite (8-22) was tested for predictability in which six variables (2-7) which yielded biographical data information were omitted as predictors. About 82 percent (.8203) of the criterion variance is estimated to be attributable to the 15 variables which essentially consist of test and major subject field data. The difference between RS value for the FM system (.8580) and the RS value for the restricted model, FM- (2-7), yields an estimate of .0377 for the unique contribution of variables (2-7) as a group. The loss in predictive accuracy when the biographical data (2-7) are not considered seems sufficient to be of practical importance. The unique contribution of the six predictor variables involving biographical data (2-7) was significant at the .01 level.

SUMMARY COMMENTS

In this study, an attempt was made to demonstrate the feasibility of utilizing Normative JAN techniques for capturing a policy that may be present. Consequently, the investigator was more interested in determining to what extent a policy had been expressed than in the actual regression equations, and the RS's only were obtained. No attempt was made to give an exhaustive analysis of the set of predictor variables. Instead, the predictor variables were grouped into three proper disjoint subsets and the unique contribution of each of these subsets was examined.

The results of this study indicate that essentially one policy was being expressed by the judges. Each of the three disjoint proper subsets contributed significantly to the prediction of the Normative JAN criterion. This suggests that this kind of admission profile data provided relevant information for the judges in the determination of their policy.

FOOTNOTES

1. The material in this article is part of a larger study submitted in partial fulfullment of the requirement for the degree of Doctor of Philosophy at Colorado State College, 1967.

2. Currently the author is a Research Specialist, Research Program, Institute for Development of Educational Activities, a Charles F. Kettering Foundation project.

REFERENCES

1. Bottenberg, R. A., and Christal, R. E., An Iterative Technique for Clustering Criteria Which Retains Optimum Predictive Efficiency, WADD-TN-61-30, AD-261 615 (Lackalnd AFB, Texas: Personnel Laboratory, Wright Air Development Division, 1961).

2. Christal, R. E., JAN: A Technique for Analyzing Individual and Group Judgment, PRL-TDR-63-3, AD-403 813 (Lackland AFB, Texas: Personnel Research Laboratory, Aerospace Medical Division, 1963).

3. Ward, J. H., Hierarchical Grouping to Maximize Payoff, WADD-TN-61-29, AD-261 750 (Lackland AFB, Texas: Personnel Laboratory, Wright Air Development Division, 1961).

4. Ward, J. H., "Multiple Linear Regression Models," Computer Applications in the Behavioral Sciences, Harold Borko (Editor) (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1962), pp. 204-37.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

THE RELATIONSHIP OF COPING STRENGTH TO SELF-CONCEPT, SCHOOL ACHIEVEMENT, AND GENERAL ANXIETY LEVEL IN SIXTH GRADE PUPILS

THOMAS M. HUGHES
Memphis State University

ABSTRACT

The primary purpose of this study was to determine the predictive quality of reading coping strength and comprehension coping strength measured under conditions of delayed auditory feedback as they relate to self-concept, anxiety, reading achievement, intelligence, and academic ability. Reading tests under delayed auditory feedback were administered to fifty-one sixth-grade students in the Memphis City School System, Memphis, Tennessee. Subjects were chosen from five elementary schools (three white and two Negro). Reading error scores, under conditions of DAF, were tabulated as a measure of reading coping strength and comprehension coping strength. Eight hypotheses were established and tested by ANOVA, t-test, correlation and multiple correlation. Overall results of the study tend to identify coping strength as a unique variable to consider in the psychology of reading as well as the psychology of such noncognitive variables as anxiety and self-concept.

The problem of this study was to make a systematic investigation of the ability of children at the sixth grade level to cope with the effects of delayed auditory feedback while reading. The ability to cope with such effects was defined as reading coping strength and comprehension coping strength. These two variables, reading coping strength and comprehension coping strength, were studied to determine their ability to predict academic success and to compare their predictive strength with that of IQ and reading achievement scores. An additional part of the problem of this study was to determine the intervariable and composite relationships between the variables of self-concept, anxiety, reading achievement, reading coping strength, comprehension coping strength, IQ and grades.

THE DISPARAGING effects of delayed auditory feedback have been substantially investigated. In a reprint of his article which appeared in a speech journal, Chase (2) listed over one hundred articles and related papers on the subject. Concurrent with the work of Chase, considerable work has been going on in delayed auditory feedback at the University of Illinois under the direction of Mowrer. Very recent research on delayed and accelerated auditory feedback has been produced at the Neurocommunications Laboratory of the John Hopkins Hospital in Baltimore, Maryland. This work is also under the direction of Chase.

When a subject speaks into a system that delays the auditory feedback of what the subject has said, it creates two distinct phenomena, i. e., increased vocal intensity and stuttering-like disfluences. For some, these effects are associated with an inability to concentrate on the speech content. The effects of delayed hearing of one's voice as he speaks were reported by Lee (5), and have since been repeated in many studies. Conversely, other subjects tend to build a gradual resistance to the disrupting effects to the point that they develop a near mastery of the phenomena.

Kerstin and Eysenck (4), in studying pursuit rotor performance as a function of different degrees of auditory distraction, came to the following conclusions: (1) distraction always lowers performance, even when the signal rate is very slow; (2) performance decrement is proportional to the amount of distraction, there being a linear relation between performance decrement and signal rate; (3) the amount of performance decrement is in part determined by individual differences, some Ss's being more distractible than others.

In the research by Baker and Madell (1) there is found the strongest rationale for continued study of the effects of distraction on certain personality variables:

> The results of the present study also illustrate clearly the principle that personality characteristics which are not operative and determinative of behavior under relatively benign circumstances may become so under more stressful conditions. This finding has important implications for improving the sensitivity of psychodiagnostic devices in general, but in particular it suggests a means of enhancing the capacity of a reading comprehension task to predict academic performance.

It is believed by this writer that a correlation may exist between the ability to cope with the effects of delayed auditory feedback while reading and academic ability. It is also believed that the variables of self-concept and anxiety level are related to reading coping strength, comprehension coping strength, and academic ability. These assumptions are supported by Spilka (7) and Rosenfeld (6).

GENERAL DESIGN

This study investigated the performance of sixth-grade Negro and white children in terms of their responses to delayed auditory feedback distraction as a measure of coping strength. Coping strength was dichotomized as reading coping strength and comprehension coping strength. Reading coping strength was determined by the child's ability to maintain vigilance during distraction by delayed auditory feedback, thus keeping his errors of omission, substitution, and pronunciation at a minimum. Comprehension coping strength was determined by the child's ability to maintain vigilance in order to retain material read under the condition of delayed auditory feedback, and to keep errors at a minimum on a test of comprehension.

This study used the process of description, prediction, and analysis. The descriptive portion established the underlying principles of feedback, self-concept, anxiety, academic achievement, and reading comprehension. The predictive portion of the study attempted to contribute to the development of a tool by which a classroom teacher, educational psychologist, or counselor may study the potential for academic ability on the basis of a child's performance on a coping strength test, his self-concept, and his level of anxiety. The analytical portion of the study utilized correlation and other statistical methods to determine the significance of differences between the performances of the groups established.

METHOD

Fifty-one sixth-grade children (26 white and 25 Negro--26 girls and 25 boys) from five elementary schools of the Memphis, Tennessee, Public School System were selected at random from names on IBM data sheets. IQ range was confined to 89-112 (stanine 4, 5, and 6) to cover average intelligence distribution, as well as to make IQ levels compatible with reading levels since it was noticed on the IBM data sheets that most IQ's below 89 and above 112 showed reading levels below or above selected reading levels of 4.0 to 8.0. These reading levels were selected to include a reasonable spread of reading abilities in sixth-grade children.

The investigator tested the children on Saturday mornings in the school cafeteria of the school which they attended. Each testing session lasted approximately 3 hours and was composed of groups of not less than seven and not more than fifteen children. All of the children were administered the Children's Manifest Anxiety Scale and the Tennessee Self-Concept Scale. Subsequently, each child was given an alternate form of the Gray, Durrell, and Spache reading tests without DAF. He was timed, scored for error, and then answered comprehension questions after the reading session. After the preliminary reading tests, the children were given a practice period on the tape recorder to become acquainted with the effects of DAF while reading. After the practice period, individual administration of the Gray, Durrell, and Spache reading tests was accomplished under DAF conditions. Each test was scored for time, errors of omission, substitution, and mispronunciation. As soon as a child finished a reading test on the tape recorder, he went to an isolated place to take the comprehension test under a proctor, and was subsequently scored for errors of comprehension. While the individual tests were being given, the other children were removed to another area so they could not hear the reading or interfere with the child being tested.

Each child's level of coping strength was determined by the total number of errors made on the three reading tests under the DAF condition and the total number of errors made on the tests of comprehension. Thus, the lower number of errors made on the reading tests indicated higher reading coping strength. In the same manner, the lower number of errors made on the comprehension tests determined higher comprehension coping strength and the higher number of errors made determined lower comprehension coping strength.

STATISTICAL PROCEDURE

Midterm grades were obtained and converted into numerical equivalents (A4, B3, C2, D1, F0). Grades were then correlated with the variables of reading achievement, IQ, self-concept, anxiety, and coping strength. Intervariable correlations were developed also and placed in a matrix. Analysis of variance and t-tests were computed to determine differences between certain variables. Multiple correlations were done to study the predictive strength of reading achievement and IQ with grades and the combination of reading coping strength and comprehension coping strength with grades. For additional purposes of comparison, a Pearson r was computed between the three reading pretests (before DAF) and grades to ascertain how effectively they predicted grades.

RESULTS

Hypothesis 1

There is no significant difference between the two coping strength scores combined and a combination of IQ and reading achievement scores in the prediction of academic grades.

The variables of reading coping strength and comprehension coping strength were placed in multiple correlation with grades as criterion. This produced a predictor r to be compared with a predictor r derived from placing the variables IQ and reading achievement in multiple correlation with grades as criterion. In Table 1 a significant multiple r of -.660 is shown for comprehension coping strength and reading coping strength with grades, with a standard error of 4.06, and a significant multiple r of .537 for IQ and reading achievement with grades with a standard error of 4.11. The negative correlations with the coping strength variables were expected. The number of errors made in DAF condition was the measure

TABLE 1

A COMPARISON OF THE MULTIPLE CORRELATION OF READING COPING STRENGTH AND COMPREHENSION COPING STRENGTH WITH GRADES AND IQ AND READING ACHIEVEMENT WITH GRADES

	r	Standard Error	z Test of Significance	P
RCS and CCS*	-.660	4.06		
IQ and Reading Achievement	.537	4.11		
			1.36	< .09

*Reading coping strength and comprehension coping strength as number of errors under DAF condition.

of coping strength. Those children with higher intelligence made fewer errors in reading and comprehension than children with lower intelligence. Therefore, children with lower IQ scores who were making poorer grades had much higher error scores than brighter children who were making good grades.

The coping strength variables yielded a somewhat greater predictor r strength than IQ and reading achievement. The z test of significance between the two correlations is shown in Table 1.

Hypothesis 1 is accepted since $p = < .05$ level of confidence was not reached. However, the .05 level of confidence was approached, and thus it appears that reading coping strength and comprehension coping strength are highly comparable with the variables of IQ and reading achievement in the prediction of grades.

Hypothesis 2

There is no significant difference between reading coping strength and comprehension coping strength in the prediction of grades.

Data in Table 2 indicated that the comprehension coping strength correlation of -.622 was higher than the reading coping strength correlation of -.498. This appears to lead to the inference that a child's ability to retain ideas or concepts while coping with the distracting effects of delayed auditory feedback is a better predictor of his academic performance than his ability simply to reduce his reading error score while coping with distraction. Apparently the ability to cope with distraction while formulating concepts in reading is more facilitating than simply inhibiting the mispronunciation and omission of words.

Of additional interest in the matrix of Table 2 is the interrelationship between the two measures of coping strength and the variables of grades, self-concept, and anxiety. The two measures of coping strength showed a significant correlation of .491 with each other. However, comprehension coping strength showed a much stronger relationship with self-concept. This may be construed to mean that children with stronger self-images were able to deal more proficiently with concepts while under distraction than to control reading error rate. Both measures of coping strength showed low and not significantly different correlations with anxiety.

Hypothesis 2 is accepted on the basis that the difference between comprehension coping strength and reading coping strength prediction of grades, as shown in Table 3, did not reach the .05 level of significance. Significance was approached with $p = < .13$.

TABLE 2

INTERCORRELATIONS OF READING COPING STRENGTH, COMPREHENSION COPING STRENGTH, GRADES, SELF-CONCEPT, AND ANXIETY

	1	2	3	4	5	6	7
1. Reading coping strength*		.491	-.498	-.182	.183	-.569	-.464
2. Comprehension coping strength*			-.622	-.479	-.111	-.483	-.334
3. Grades				.411	-.108	.527	.412
4. Self-concept					-.462	.099	.113
5. Anxiety						.064	.165
6. IQ							.641
7. Reading achievement							

*Coping strength was the number of error.

Hypothesis 3

There is no significant difference between self-concept scores of those children who demonstrate high coping strength and those who demonstrate low coping strength.

It was originally anticipated that children with high coping strength would show high self-concept scores, and children with low coping strength would show low self-concept scores. No significant difference existed in the self-concept scores of the high and low coping strength groups. Reading coping strength was the lesser of the two coping strengths in the approach to significance. The self-concept scores in the high comprehension coping strength group approached, but did not attain, the .05 level of significance. Hypothesis 3 is accepted because statistical significance was not obtained.

TABLE 3

TEST OF SIGNIFICANCE BETWEEN NON-INDEPENDENT CORRELATION COEFFICIENTS

	CCS*	RCS**	RCS with CCS	Significance	p
r	-.622	-.498	.491	t = -1.13	< .13

*Comprehension coping strength with grades.
**Reading coping strength with grades.

Hypothesis 4

There is no significant difference in grades between high and low self-concept groups and high and low IQ groups nor any interaction between self-concept and IQ.

Analysis of variance data are presented in Table 4 which supports rejection of Hypothesis 4. The high IQ group earned significantly higher grades than the low IQ group. In addition children with a high self-concept tended to earn higher grades than children with low self-concepts. Interaction effects were not observed.

Hypothesis 5

There is no significant differences between the anxiety levels of low and high IQ groups.

Hypothesis 5 is accepted. Data from t-tests indicated children with high IQ's showed a slightly higher but not statistically different level of anxiety when compared to the low IQ group.

Hypothesis 6

There is no significant difference between the reading coping strength and comprehension coping strength scores of those children who have IQ scores above the mean and those who have IQ scores below the mean.

Hypothesis 6 is rejected. Children with IQ measures above the mean of 101.56 showed a significantly lower number of errors when compared with low IQ group. This may be interpreted to mean that children with higher IQ's may be expected to cope better with distraction and to make significantly fewer errors in reading and comprehension under the condition of delayed auditory feedback.

TABLE 4

ANALYSIS OF VARIANCE OF GRADES BY IQ AND SELF-CONCEPT (MEAN SPLIT OF GROUPS)

Source	SS	df	MS	F	P
IQ	186.06	1	186.06	7.47	< .01
Self-concept	96.24	1	96.24	3.86	< .10
IQ x SC	18.74	1	18.74	.75	
Error	1171.08	47	24.92		
Total	1472.12	50			

Hypothesis 7

There is no significant difference between Negro and white students on measures of coping strength, anxiety and self-concept.

Hypothesis 7 is rejected because reading coping strength of Negro pupils was statistically superior.

Significant differences did not exist between Negro and white students on measures of the variables of self-concept and anxiety. The white students showed less anxiety and more positive self-concepts, but the level did not approach significance.

Hypothesis 8

There is no significant differences between boys and girls on measures of coping strength, anxiety, and self-concept.

Hypothesis 8 is accepted. Reading coping strength did not vary significantly between boys and girls. Self-concepts did not vary significantly between boys and girls although girls had a slightly higher mean self-concept. Boys had greater comprehension coping strength than girls. The anxiety level of girls was higher than boys with the difference approaching significance.

PREDICTING GRADES FROM COPING STRENGTH

As previously stated, part of the problem of this study was to determine the predictive quality of coping strength. A predictor formula by Garrett (3) is shown as follows to predict a composite total of eight grades which are given in elementary school:

$$\text{Y pred.} = r\frac{\sigma y}{\sigma x} \cdot \;\; X \;\_\_\_\; r\,\frac{\sigma y}{\sigma x} \cdot \;\; Mx + My.$$

r = correlation between X and Y.
X = reading coping strength score.
Y = numerical equivalent of eight subject grades.

σ y and σ x = the standard deviations of the X and Y scores.
Mx and My = the means of the X and Y scores.
Y pred. = the predicted score of grades.

Computing this formula we find:

$$\text{Y pred.} = \frac{-.50 \times 5.477}{14.036}\, X \;\_\_\; \frac{-.50 \times 5.477 \times 22.78 + 18.35}{14.036}$$

and

$$\text{Y pred.} = .19X + 13.91.$$

If, for example, a student made seven (number of errors) on the reading coping strength test, his most likely composite of grade total would equal 32 or A average. This is accomplished by subtracting the number of errors he made from 100, then applying the predictor formula.

$$\begin{array}{l} 100 - 7 = \;\; .93 \\ \qquad\qquad \underline{\;\;.19} \\ \qquad\quad \overline{17.67} + 13.91 = 31.58 \end{array}$$

From a composite of eight grades on the five point system (A=4, B=3, C=2, D=1, F=0) it may be seen that a student with a reading coping strength of seven should have a grade average of A. The lower the reading coping strength error score, the higher we should predict his grades. The same would hold true for comprehension coping strength.

In the application of such a formula, it should be understood that grade meaning varies among schools. An A in one school does not necessarily mean that the same standards were met in another school by a student who made an A. Therefore, should a coping strength score be used for prediction, the standards of the individual school would have to be taken into consideration.

PREDICTING GRADES FROM READING TESTS

Before each child was administered the coping strength tests, he was given three timed paragraphs to read. These paragraphs were from the Spache, Durrell, and Gray tests and were administered free of the conditions of delayed auditory feedback in order to observe how well they predicted grades in comparison with material read under the conditions of delayed auditory feedback. Table 5 shows that the error scores on the combined tests predicted grades at -. 54r and the comprehension portion predicted at -.31r. When these results are compared to the correlations in Table 2, it is seen that the reading tests accomplished without delayed auditory feedback had a slightly higher r than those under the DAF conditions (-. 498). However, the comprehension tests under DAF (-. 622) showed a much stronger correlation than the -31r of tests without DAF. These results appear to point out that children who can maintain greater viligance or attention under the condition of DAF retain more of the material read, although reading error rate does not vary significantly from one level of vigilance to another.

In general, the results of the study show that coping strength is an important variable to consider in the study of individual differences, in reading ability, and in the prediction of academic success.

TABLE 5

PEARSON r OF THREE READING TESTS
(BEFORE DAF) AND GRADES

	Reading	Comprehension
r	-. 54	-. 31

DISCUSSION

The overall results of this study tend to identify coping strength as a unique variable to consider in the psychology of reading as well as the psychology of such noncognitive variables as anxiety and self-concept. Individual differences were shown in the demonstration of coping strength as well as the other variables studied.

One of the most interesting observations made by this investigator, while working with the children in the sample, was their exceptional motivation to continue the exercise of reading with delayed auditory feedback long after their testing period was over. Their enthusiasm seemed to suggest that in the act of reading, the introduction of this type of distraction (DAF) called forth a vigilance that was actually enjoyable. It seems reasonable to assume that in addition to the enjoyment of performing in such a manner, the children found it highly motivating to be involved in a feedback loop that was individualized according to the pattern of their own performance.

Psychologists, who study individual differences under laboratory conditions, and educators, who teach them in the classroom, should consider not only the prediction aspect of coping strength, but also the practical value of giving children ample opportunity to participate in it so that it may sharpen attention, set, and vigilance skills, and thereby improve the art of reading and the science of information processing in students. For the psychologist, the implications of continued research into the dimensions of coping strength seem highly warranted.

REFERENCES

1. Baker, R. W., and Madell, T. Q, " A Continued Investigation of Susceptibility to Distraction in Academically Underachieving and Achieving Male College Students," Journal of Educational Psychology., LVI (1965), 257 pp.
2. Chase, R. A., Sutton, S., and First, Daphne. Bibliography: Delayed Auditory Feedback," Journal of Speech and Hearing Research, II (1959), pp. 193-200.
3. Garrett, H. E. Elementary Statistics (New York: David McKay Co., 1964), 101 pp.
4. Kerstin, S., and Eysenck, H. J. "Pursuit Rotor Performance as a Function of Different Degrees of Distraction," Life Science, IV, (1965), pp. 889-897.
5. Lee, B. S. "Effects of Delayed Speech Feedback," Journal of Acoustical Society of America, XXII (1950), pp. 824-826.
6. Rosenfeld, J. M. "Some Perceptual and Cognitive Correlates of the High Need for Social Approval Personality," Dissert. Abstr., XXIV (1964), 4805 pp.
7. Spilka, B. "Relationships Between Certain Aspects of Personality and Some Vocal Effects of Delayed Speech Feedback," Journal of Speech and Hearing Disability, XIX (1954), pp. 491-503.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

THE ANALYSIS OF MEANS TECHNIQUES: THEORY AND APPLICATION[1]

SHAWKY F. KARAS

Southern Connecticut State College

ABSTRACT

The theoretical basis for the Analysis of Means Technique is discussed. In addition, a simplified working procedure is outlined, step-by-step for an actual problem. The data for the problem are analyzed by the analysis of means techniques, which compares differences between means instead of population variances estimates. The data are also analyzed by the analysis of variance technique. The conclusions reached by both techniques are the same. The graphical representation of the ANOM is helpful in understanding data. The ANOM technique is usually used in conjunction with the analysis of variance, either as an initial analysis of data or to augment the analysis of variance.

THE THEORECTICAL basis for the Analysis of Means Technique is presented for the first time in an educational journal. In addition, a simplified working procedure is outlined, step-by-step for an actual problem. The data for the problem are analyzed by the analysis of means technique, which compares differences between means instead of population variance estimates.

The analysis of means technique is an extension of the Shewhart Control chart (1) which is used in industry for quality control. The analysis of means is well known in industry but it is not generally known among psychologists and educators. The method is useful because of its simplicity and comparability to the analysis of variance.

The procedure tests the null-hypothesis that k group means represent random samples drawn from a universe. If the null-hypothesis is rejected with risk α, there are significant differences among the group means. If there are only two means, the analysis is the student– test drawn as a graph. If $k > 2$, the null-hypothesis of randomness will be rejected with risk α, if at least one mean lies outside the two decision lines, drawn for the graphical analysis of means of the experiment.

The graphical representation of the analysis of means is helpful in understanding data. It provides a comparison of magnitudes of differences along with a test of statistical significance. It can pinpoint the presence of certain types of effects which may be missed by other techniques. The ANOM technique is usually used in conjunction with the analysis of variance, either as an initial analysis of the data or to augment the analysis of variance. The method is outlined in the following paragraphs, followed by an example which was analyzed using both the analysis of variance and analysis of means methods.

THE ANALYSIS OF MEANS METHOD

1. Standard Deviation Given. Let us assume a population with mean $\overline{X'}$ and standard deviation σ'. If we have k independent samples of n each from the given population and we want to consider all the k means simultaneously, we compute the interval.

$$\overline{x'} \pm z_{\alpha} \sigma'_{\overline{x}}$$

The k means should lie within this interval to accept the null-hypothesis with risk α. In order to compute $Z\alpha$, let p represent the probability that one mean will lie between the two decision lines drawn at $\bar{X}' - Z_{\alpha} \sigma_{\bar{x}}$ and $X' + Z_{\alpha} \sigma'_{\bar{x}}$. Then the probability that all k means will lie within these lines is p^k. If at least one point lies outside these decision lines, we reject the null-hypothesis. If we consider the type I error as α (the risk) then Z_{α} may be computed by the following steps:

$$p^k = 1 - \alpha$$
$$k \log p = \log (1 - \alpha)$$

From the table of areas under the normal curve we obtain Z_{α} corresponding to the value of p such that the area between the ordinates drawn at $\bar{X}' - Z_{\alpha}\sigma' \bar{x}$ and $\bar{X}' + Z_{\alpha} \sigma'_{\bar{x}}$ is p. Clearly Z_{α} is a function of k and α, so it is defined as Z_{α} (k).

2. No Population Standard Deviation Given. If $k = 2$ the method is similar to the Student's t-test of testing the significance of the difference between two means. When $k > 2$ the appropriate decision lines depend upon the degrees of freedom (ν) in the estimate $\hat{\sigma}$. If k means, each based on n observations, are to be analyzed simultaneously for non-randomness the decision lines will be drawn at

$$\bar{\bar{X}} + H_{\alpha}(k, \nu).$$

$\bar{\bar{X}}$ is the grand mean (mean of means). If any one of the means lies outside the decision lines, the null-hypothesis will be rejected and this will be an indicate of non-randomness. It is clear that H is a function of k, α, and ν. There are two sets of tables which locate these decision lines. One set is by Halperin, Greenhouse, Cornfield, and Talokar (2) and the other by Nair (3). Ott (4) uses the first set of tables with some modifications to compute $H_{\alpha}(k, \nu)$ for some k and combinations with $\alpha = .05$ and $\alpha = .01$. The tables are reproduced as Table 3 and 4 in the appendix.[2]

Illustrative Example

As a part of an experimental study, the College Entrance Examination Board's Scholastic Aptitude Test (SAT) was administered to students in the Fifth Form (roughly equivalent to the twelfth grade in the U. S. A.) and the Sixth Form (roughly equivalent to the Junior College in the U. S. A.) in England. The SAT consists of a verbal section and a mathematical section. The tests were administered to male and female students in single sexed schools in Northern and Southern England. The data are reported in Table 1 for the verbal section of SAT. Table 2 shows the analysis of variance using a three way factorial design. The analysis of means will be illustrated in the following steps:

The Analysis of Means

1. For Main Effect

If we have two or more independent variables

g_1------g_{n1}
h_1------h_{n2}
w_1------w_{n3}

We find the sample means for each variable

$\bar{X}_{g_1}$ ----- $\bar{X}_{g_{n_1}}$

$\bar{X}_{h_1}$ -----$\bar{X}_{h_{n2}}$

$\bar{X}_{w_1}$ -----$\bar{X}_{w_{n3}}$

We plot the $\bar{X}_{g1}$------$\bar{X}_{gn3}$ values on one chart, and the $\bar{X}_{h1}$------$\bar{X}_{hn2}$ values on the same chart extended ($\bar{\bar{X}}$ is the grand mean). We also determine for each of the different independent variables g, h, etc. the number of subgroups (of n each) which are pooled in comparing differences in the effects of g_1-----g_{n1} and subsequently the effects of h_1-----h_{n2}, etc. Then we compute the standard deviations $\sigma_{\bar{x}_g}$, $\sigma_{\bar{x}_h}$, ----, etc. We find H_{05} and H_{01} from Tables 3 and 4. We draw two sets of decision lines at $\bar{\bar{X}} \pm H_{05}\hat{\sigma}_{\bar{x}g}$ and $\bar{\bar{X}} \pm H_{01}\hat{\sigma}_{\bar{x}g}$ for the g variables and similar decision lines for the other variables.

TABLE 1

THE THREE-WAY DESIGN OF SAT-V IN ENGLAND *

	North		South	
	Boys	Girls	Boys	Girls
Fifth	N = 69 Sum = 31102 Mean = 450.7536 S. D. = 85.0753	N = 69 Sum = 34602 Mean = 501.4783 S. D. = 63. 1002	N = 69 Sum = 30933 Mean = 448.3043 S.D. = 79.3461	N = 69 Sum = 32726 Mean = 474.2899 S.D. = 70.9849
Sixth	N = 69 Sum = 37835 Mean = 548. 3333 S.D. = 81.9420	N = 69 Sum = 38447 Mean = 557.2029 S.D. = 66.0335	N = 69 Sum = 38514 Mean = 558.1739 S.D. = 71.0406	N = 69 Sum = 36878 Mean = 534.4638 S.D. = 76.2511

*The writer wishes to express his appreciation to Mr. John Fremer at Educational Testing Service for providing him with the data which were collected for a study now in progress.

TABLE 2
THE ANALYSIS OF VARIANCE

S. V.	D. F.	S. S.	M. S. S.	E(M. S. S.)	F
Total	551	4079284.38			
Forms (F)	1	901776.57	901776.57	$\sigma^2 + 276\,\sigma_F^2$	162.153**
Sex (S)	1	33015.15	33015.15	$\sigma^2 + 276\,\sigma_S^2$	5.936*
Region (R)	1	15605.53	15605.53	$\sigma^2 + 276\,\sigma_R^2$	2.806
F X R	1	2416.77	2416.77	$\sigma^2 + 138\,\sigma_{FXR}^2$	0.434
F X S	1	72290.84	72290.84	$\sigma^2 + 138\,\sigma_{FXS}^2$	12.999**
R X S	1	28336.95	28336.95	$\sigma^2 + 138\,\sigma_{RXS}^2$	5.095*
F X R X S	1	530.16	530.16	$\sigma^2 + 69\,\sigma_{FXRXS}^2$	0.095
Residuals	544	3025312.41	5561.24		

*$p < .05$
**$p < .01$

$\hat{\sigma} = \sqrt{5561.24} = 74.57$

The Upper Decision Line (UDL) with (.05)
$= 509.12 + 1.39 \dfrac{(74.57)}{\sqrt{276}}$
$= 509.12 + 6.24$
$= 515.36$

The Lower Decision Line (LDL) with (.05)
$= 509.12 - 6.24$
$= 502.88$

The UDL with (.01) $= 509.12 + 1.83 \dfrac{(74.57)}{\sqrt{276}}$
$= 509.12 + 8.21$
$= 517.33$

The LDL (.01) $= 509.12 - 8.21$
$= 500.91$

2. For Two-Factor Interaction
(a) Form and Region

	North	South
Fifth	476.12	461.30
Sixth	552.77	546.32
Difference	76.65	85.02

$\bar{\bar{\triangle}}$ (Difference Mean) $= 80.83$

UDL (.05) $= 80.83 + 1.39 \sqrt{2} \dfrac{(74.57)}{\sqrt{138}}$
$= 80.83 + 12.48$
$= 93.31$
LDL (.05) $= 80.83 - 12.48$
$= 68.35$
UDL (.01) $= 80.83 + 1.83 \dfrac{(74.57)}{\sqrt{138}} \sqrt{2}$
$= 80.83 + 16.48$
$= 97.26$
LDL (.01) $= 80.83 - 16.48$
$= 64.40$

(b) Form and Sex

	Fifth	Sixth
Boys	449.53	553.25
Girls	487.88	545.83
	38.35	- 7.42

$\bar{\bar{\triangle}} = 15.47$

UDL (.05) $= 15.47 + 12.48 = 27.95$
LDL (.05) $= 15.47 - 12.48 = 2.99$
UDL (.01) $= 15.47 + 16.43 = 31.90$
LDL (.01) $= 15.47 - 16.43 = -.96$

(c) Region and Sex

	North	South
Boys	499.54	503.24
Girls	529.34	504.38
	30.39	1.14

$\bar{\bar{\triangle}} = 15.76$.

UDL (.05) $= 15.76 + 12.48 = 28.34$
LDL (.05) $= 15.76 - 12.48 = 3.28$
UDL (.01) $= 15.76 + 16.43 = 32.19$
LDL (.01) $= 15.76 - 16.43 = -.67$

TABLE 3
PERCENTAGE POINTS FOR THE STUDENTIZED MAXIMUM ABSOLUTE DEVIATE IN NORMAL SAMPLES

$\alpha = .05$ $\quad H = \left[\text{Max} \quad \frac{(\bar{X}_n - \bar{\bar{X}})}{\hat{\sigma}\bar{x}} \; , \; \frac{(\bar{\bar{X}} - \bar{X}_1)}{\hat{\sigma}\bar{x}} \right]$ No standard deviation given

df \ k =	2	3	4	5	6	8	10	15	20	30
5	1.82									
6	1.73	2.59	2.94	3.19	3.37					
8	1.63	2.39	2.71	2.92	3.09	3.33				
10	1.58	2.29	2.58	2.78	2.93	3.15	3.31			
15	1.51	2.16	2.42	2.60	2.74	2.93	3.07	3.32		
20	1.48	2.10	2.35	2.52	2.64	2.85	2.96	3.18	3.33	
30	1.44	2.04	2.28	2.44	2.56	2.73	2.86	3.06	3.19	3.37
40	1.43	2.01	2.25	2.40	2.52	2.69	2.80	3.00	3.13	3.29
60	1.41	1.98	2.21	2.36	2.48	2.64	2.76	2.94	3.06	3.22
120	1.40	1.95	2.18	2.33	2.44	2.60	2.71	2.88	3.00	3.15
	1.39	1.93	2.15	2.29	2.40	2.55	2.65	2.82	2.94	3.08

k = number of means being compared

df = degrees of freedom in estimate of error variance

The values in this table are the averages of the upper and lower bounds given by Halperin, Greenhouse, Cornfield and Zalokar.

In the special case where we have 2^3 factorial design the following procedures can be used:

(a) Form and Region

	North	South
Fifth	476.12	461.30
Sixth	552.77	546.32

Mean of the Fifth and North (FN) and Sixth and South (SS) = 511.22
Mean FS and SN = 507.04

(b) Form and Sex

Mean BF and GS = 497.68
Mean BS and GF = 520.56

(c) Region and Sex

Mean BN and GS = 501.96
Mean BS and GN = 516.29

The two decision lines for each two-factor interaction are identical with that for the main effects.

3. For Three-Factor Interaction

Mean NBF, NGS, SBS, and SGF = 510.10
Mean NGF, NBS, SBF, and SGS = 508.14

NBF: The cell which is the combination of North, Boys, and Fifth Form. The two decision lines for the three-factor interaction are identical with that for the main effects.

Comparing the Results of the Analysis of Variance and the Analysis of Means

Table 2 shows the results of the analysis of variance: Figure 1 shows the results of the analysis of means. From Table 2 the main effect of Forms (F), and the Form by Sex interaction are significant at the .01 level; the main effect of Sex, and the Region by Sex interaction are significant at the .05 level. From Figure 1 we make the same conclusions. Figure 2

TABLE 4
PERCENTAGE POINTS FOR THE STUDENTIZED MAXIMUM ABSOLUTE DEVIATE IN NORMAL SAMPLES

$\alpha = .01$ $H = \left[\text{Max} \left(\frac{\bar{X}_n - \bar{\bar{X}}}{\hat{\sigma}_{\bar{X}}}, \frac{\bar{\bar{X}} - \bar{X}_1}{\hat{\sigma}_{\bar{X}}} \right) \right]$

No standard deviation given

df \ k =	2	3	4	5	6	8	10	15	20	30	40	60
5	2.85											
6	2.62	3.74	4.21	4.53	4.78							
8	2.37	3.31	3.70	4.97	4.17	4.47						
10	2.24	3.08	3.43	3.67	3.86	4.11	4.29					
15	2.08	2.81	3.12	3.32	3.47	3.69	3.84	4.11				
20	2.01	2.70	3.98	3.17	3.30	3.50	3.63	3.87	4.02			
30	1.94	2.58	2.85	3.02	3.15	3.33	3.45	3.66	3.79	3.96		
40	1.91	2.53	2.79	2.95	3.07	3.24	3.36	3.56	3.68	3.84	3.96	
60	1.88	2.48	2.73	2.88	3.00	3.16	3.27	3.46	3.58	3.73	3.84	3.97
120	1.85	2.43	2.67	2.82	2.93	3.09	3.20	3.37	3.48	3.62	3.72	3.86
Inf.	1.82	2.39	2.61	2.76	2.87	3.02	3.12	3.29	3.39	3.53	3.62	3.73

k = number of means being compared.
df = degrees of freedom in estimate of error variance

The values in this table are the average of the upper and lower bounds given by Halperin, Greenhouse, Cornfield and Zalokar.

is helpful in understanding the first-order interactions. It is clear that the Form by Region interaction is not significant because the two lines are approximately parallel while the two lines of the other two interactions are not parallel which indicates that their interactions are significant.

FOOTNOTES

1. This method was developed by Professor Ellis Ott, Director of the Statistics Center at Rutgers, the State University, New Brunswick, New Jersey.

2. Tables 3 and 4 are reproduced by the permission of Professor Ott.

3. The standard deviation estimate (74.57) is taken from Table 2 or we can obtain an independent estimate.

REFERENCES

1. Grant, E. L., Statistical Quality Control (New York: McGraw Hill, 1964).

2. Halperin, M., Greenhouse, S. W., Cornfield, J., and Zalokar, J. "Tables of Percentage Points for the Studentized Maximum Absolute Deviate in Normal Sample," Journal of American Statistical Association, L (1955), pp. 185-195.

3. Nair, K. R., "The Distribution of the Extreme Deviate from the Sample Mean and its Studentized Form," Biometrika, XXXV (1948) pp. 118-144.

4. Ott, E. R., Analysis of Means, Technical Report No. 1, 1958. Prepared for Army, Navy, and Air Force under Contract No. 404 (11) (Task NR 042 - 021) with the Office of Naval Research.

FIGURE 1

THE ANALYSIS OF MEANS

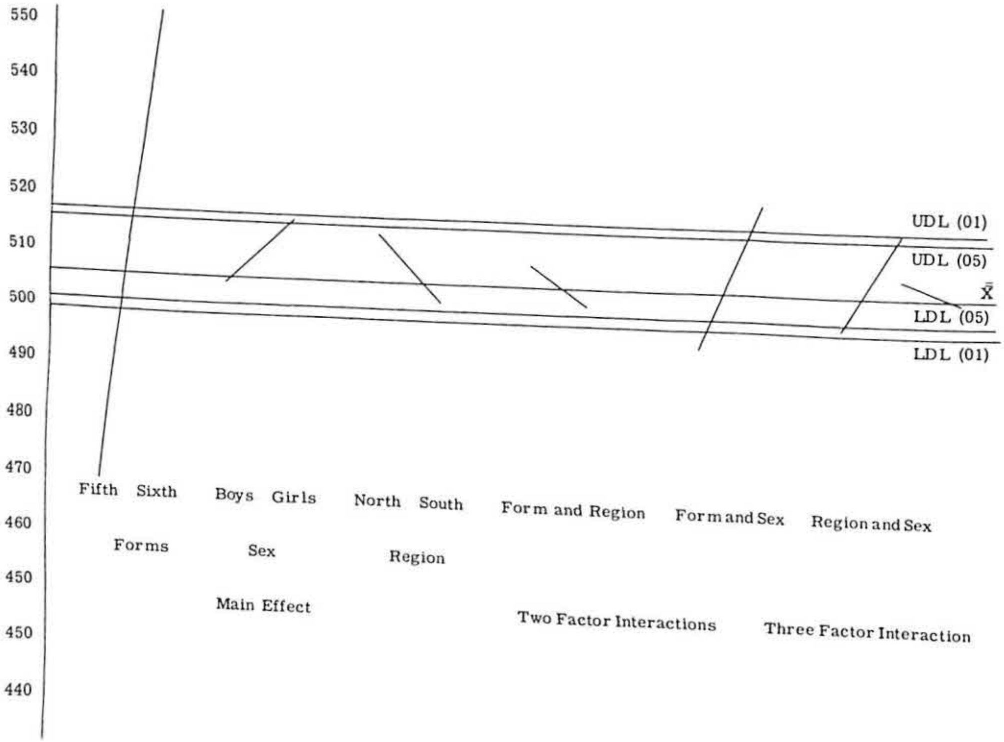
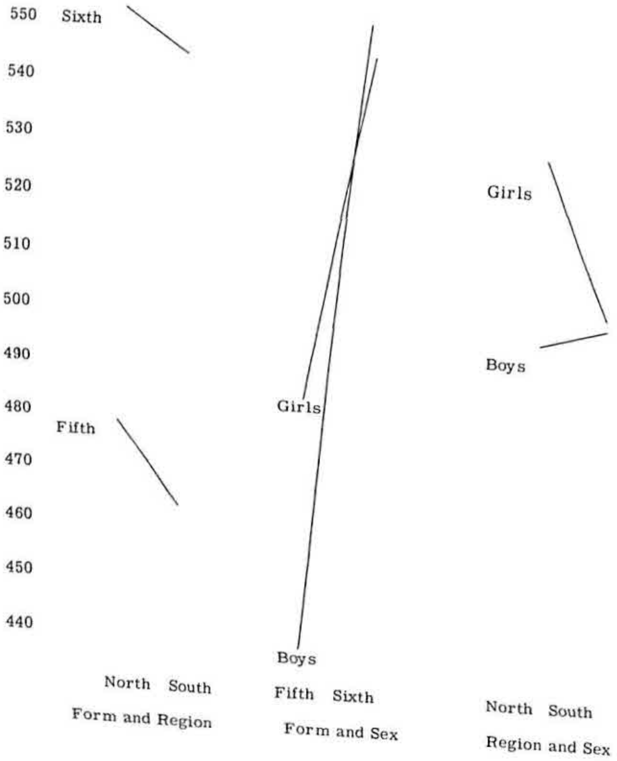


FIGURE 2

FIRST ORDER INTERACTIONS



THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

BAYESIAN INFERENCE

JON C. MARSHALL

University of Missouri - St. Louis

ABSTRACT

A practical situation is used to illustrate how Bayesian inferential procedures can be applied to educational problems. In a comparison of Bayesian and classical approaches to the same problem, it is shown that the formal process involved in Bayesian procedure is different from and more objective than classical procedure.

The purpose of this paper is to provide a clear illustration of the use of Bayesian inferential procedures. To this end, an educational problem is presented and a step-by-step Bayesian solution is provided. It is illustrated that the Bayesians have come up with a cogent technique in situations where the decision-maker is faced with choosing between two alternatives to a problem, one of which must be accepted.

THE BAYESIAN approach to statistics has been both damned and praised by behavioral researchers. Research theorists seem to present Bayesian and Classical Methods as two mutually exclusive research philosophies. Yet the Bayesian philosophy offers definite advantages to the researcher in many situations, particularly in instances when a direction of action will be decided upon at the completion of the research. In other situations, the classical approach seems to hold the edge. Classical methods are more amenable to most complex problems involving multiple treatments and/or multivariate data.

A recently published article by Marshall (1) illustrated how Bayesian decision theory could be applied to counseling and educational problems involving classical regression analysis. The purpose of this paper is to present a practical situation which will illustrate how Bayesian statistical theory can be applied to behavioral problems involving classical hypothesis testing.

PROBLEM

The chairman of the mathematics department at Hill Top High School felt that "modern" mathematics should be introduced into the curriculum. The city superintendent, who was somewhat pessimistic about such a change, was willing to consider the idea only if it could be shown that it was superior to the traditional method. Since the cost of new supplies and the retraining of teachers is high, a full scale experimental program was impossible. But, it was possible to take a random sample of students (for expedience, say ten), give them a "modern" mathematics program, and then compare them with students having taken a traditional program.

CLASSICAL APPROACH

The classical researcher would typically draw two samples of students, give one sample the "modern" mathematics treatment, and the other the traditional mathematics treatment. Each group would be given the same appropriate achievement test after the treatments and then a one-tailed test (for example the t-test) would be used to test the null hypothesis, $H_0: \mu_1 - \mu_2 = 0$. The alternative would be $H_1: \mu_1 > \mu_2$. Thus, the null hypothesis would be retained if $\mu_1 - \mu_2 < K\alpha$ and rejected if $\mu_1 - \mu_2 \geq K\alpha$, where $K\alpha$ is the appropriate value of the statistic for rejection of the null hypothesis at α level of significance. Thus, the final decision may be made in terms of this acceptance or rejection of the null hypothesis. If the null hypothesis is retained, the request to change the program would be denied; while, if it is rejected the request to change the program would be at least considered and possibly adopted. It is often argued that this is an objective decision. But in reality, the researcher has to make several subjective decisions before the final product is seen. First, he has to decide on the level of significance (α). This should be a function of the criticalness of making a correct decision, but it is often decided by default by rigidly placing $\alpha = .05$. Thus, if done properly, the determination of α is a subjective decision. Secondly, what about the size of N? This value is determined by taking as large a sample as possible or by deciding what magnitude of difference between the means will be considered practically significant. The final decision of the school superintendent is again subjective. In light of the evidence, he still has to decided whether the practicality of the difference is worth the cost, a subjective decision which might be reflected in the sample size. Thus, in at least three places in the decision making process the classical researcher has to subjectively appraise the situation. The Bayesian recognizing this subjectivity tries to formulate it into more precise terms.

BAYESIAN APPROACH

Introduction

The approach outlined here will differ from the classical approach on at least three accounts: 1) α

will be of little importance, 2) the criticalness of making a correct decision will be formally stated in terms of the loss function, and 3) the analysis will be made using the a posterior distribution rather than the a priori distribution. Furthermore, the hypotheses will be stated as the proportion of the "modern" mathematics treatment group above or below a given point rather than the magnitude of difference between the two means, an approach which is open to the classical researcher.

Hypotheses

The mean score on an appropriate standardized achievement test for those students in the traditional program was available from previous classes. Hence, the superintendent formulated the problem in terms of θ, the proportion of students in the "modern" mathematics treatment group above the mean (M) of the traditional group. Thus, the hypotheses can be stated as follows:

$$H_0: \theta \leq .50$$
$$H_1: \theta > .50$$

Bayes's Theorem

Before continuing with the problem, let us consider Bayes's Theorem.

$$P(E_i \mid A) = \frac{P(E_k) \times P(A \mid E_k)}{\sum_{i=1}^{j} P(E_i) \times P(A \mid E_i)} ;$$

where $P(A) \neq 0$; $i = 1, 2, \ldots, j$; E_i represents the various values of θ; and A is a particular experimental value that has been observed. In other words,

$$\left[\begin{matrix}\text{probability of}\\ \text{having a specific}\\ \text{state of nature}\\ \text{given an experi-}\\ \text{mental result}\end{matrix}\right] = \frac{\left[\begin{matrix}\underline{\text{a priori}}\text{ probability}\\ \text{of the state of nature}\end{matrix}\right] \times \left[\begin{matrix}\text{Probability of obtaining}\\ \text{the experimental result}\\ \text{given the state of nature}\end{matrix}\right]}{\sum\limits_{\substack{\text{for all}\\ \text{possible}\\ \text{states of}\\ \text{nature}}} \left[\begin{matrix}\underline{\text{a priori}}\text{ proba-}\\ \text{bility of the}\\ \text{state of nature}\end{matrix}\right] \times \left[\begin{matrix}\text{Probability of obtaining}\\ \text{the experimental result}\\ \text{given the state of nature}\end{matrix}\right]}$$

The state of nature in this case refers to the actual proportion of a population of "modern" mathematics students who would score above the mean score for the "traditional" students (previously noted as θ). From the Bayesian viewpoint, the purpose of an experiment is to modify one's opinion rather than discover something new. That is, there is a non-zero probability of occurance which can be personally assigned to any event. The purpose of an experiment is to modify this personal probability in light of new evidence. Therefore the a priori probabilities corresponding to the various states of nature $[P(E_i)]$ have to be determined as an initial stage in the experimental process. The experimental probability $[P(A \mid E_i)]$ is simply the binomial probability of obtaining the experimental value of A if the population value is E_i.

A Priori Distribution

Knowing very little about the actual distribution of scores, the superintendent decided that it was just as likely that 50 percent of the sample would score above the mean of the traditional group as it was that 45 percent or 55 percent would score above the mean. It was further argued by the superintendent that more than 60 percent of a "modern" mathematics population would not score above the mean of the traditional group or the value of "modern" mathematics would be so obvious as to warrant any other consideration impossible and that if less than 40 percent would score above the mean, the ineffectiveness of the program would be so obvious as to have no supporters. Thus, it was decided that the a priori distribution should be rectangular with a total range of 21 points, with end points at .40 and .60. Therefore, it can be determined that the $P(E_i) = 1/21 = .048$, where $E_1 = .40$, $E_2 = .41$, $\ldots$, $E_{21} = .60$ (See Figure 1).

It should be noted that the distribution with which we are working is binomial. That is, there are only two possible observations that can be made. The first is if the observed proportion of scores (A/n) above the value M is $> .50$ and the second is if the observed porportion of scores above the value M is $\leq .50$. Thus, the probability function is

$$P(A = a) = \begin{cases} .48 \text{ for } A/n > .50 \\ .52 \text{ for } A/n \geq .50 \end{cases}$$

That is, the a priori probability of $A/n > .50$ is .48. This can be determined by adding the a priori probabilities $[P(E_i)]$ for $E_i = .51, .52, \ldots, .60$.

Loss Ratio

Before continuing, the superintendent had to decide what probabilities that $\theta > .50$ (H_1) would be sufficient for him to take action. This decision is dependent upon the losses of utility. That is, what would be his loss if he decided $\theta \leq .50$ when $\theta > .50$, $L(\theta_0)$, and what would be his loss if he decided $\theta > .50$ when $\theta \leq .50$, $L(\theta_1)$. These loss values are determined using both objective and subjective criteria such as the amount of expenditures, loss to students taught using an inferior method, loss of reputation of the superintendent for making a bad decision, and loss of time spent. The losses might be figured as follows:

I. Possible loss when deciding in favor of "modern" mathematics (H_1)

A. Expenditures	
1. New "modern" materials - \$5,000	
2. Retraining of teachers - \$5,000	10 units
B. Students	5 units
C. Reputation	2 units
D. Time	5 units
$L(\theta_1)$ Total	22 units

II. Possible loss when deciding in favor of traditional mathematics (H_0)

A. Expenditures - "Up-to-date" material - \$2,000	2 units
B. Students	5 units
C. Reputation	2 units
D. Time	2 units
$L(\theta_0)$ Total	11 units

The relative weight of the importance of each loss is strictly personal to the decision-maker. It could be the importance of each type of error as he sees it or it could be the importance of each as a committee sees it. But, it is a personal weighting of each factor. For purposes of simplicity let us assume that the weighting was personal to the superintendent. Then, as he saw it, the loss of \$2,000 was as important as the loss of reputation he would incur if he was wrong. The loss to the students would be 2 1/2 times greater than his personal loss. There was a time loss in both cases, but the time loss when deciding in favor of the change was much greater than the time loss when deciding against the change. After expressing each segment in relative units, the relative loss for each error can be determined. The relative losses are thus $L(\theta_0) = 11$ and $L(\theta_1) = 22$.

The possible loss when deciding in favor of H_1 is twice as great as when deciding in favor of H_0. This is expressed as a decimal fraction in the loss ratio. The loss ratio can be defined as:

$$L(\theta) = \frac{L(\theta_0)}{L(\theta_1)} \quad \frac{11}{22} = 1/2.$$

This loss ratio will play a key role in the final decision.

A Posteriori Distribution

Suppose that the sample of ten students was treated using "modern" mathematics. What is the probability of a state of nature existing (E_i) if a certain experimental value (A) is observed (i.e., $P(E_i \mid A)$)? This is nothing more than calculating the probability for each E_i using Bayes's Theorem. But, to do this the probability for A given E_i (i.e., $P(A \mid E_i)$) is needed. Since the prior distribution is rectangular and thus $P(E_1) = P(E_2) = \ldots = P(E_i)$, .048, the $P(E_i)$ term can be factored out of both the numerator and denominator of Bayes's theorem, or

$$P(E_k \mid A) = \frac{P(E_k)\, P(A \mid E_k)}{\Sigma P(E_i)\, P(A \mid E_i)} = \frac{P(E_i)\, P(A \mid E_k)}{P(E_i)\, \Sigma P(A \mid E_i)} = \frac{P(A \mid E_k)}{\Sigma P(A \mid E_i)}.$$

The prior conditional probabilities $P(A \mid E_i)$ for A = 2, 3, ..., 8 given E_i are shown in Table 1 along with the accumulated posterior probabilities $P(E_i \mid A)$ for E_i given A = 2, 3, ..., 8. The graphs of these accumulated posterior probabilities are also shown in Figure 1. The prior conditional probabilities $[\,P(A \mid E_i)\,]$ are nothing more than those probabilities determined using the binomial distribution where n = 10, x = A, and p = E_i.

$$P(A \mid E_i) = \binom{n}{A} (E_i)^A \, (1 - E_i)^{n-A}.$$

The posterior probabilities are those determined through the use of Bayes's theorem. For example,

$$P(A = 2 \mid E_i = .40) = \binom{10}{2} (.40)^2 \, (.60)^8 = .121 \text{ and}$$

$\sum_{i=1}^{21} P(A = 2 \mid E_i) = 1.092$. Therefore, the posterior probability for $E_i = .40$ is

$$P(E_i = .40 \mid A = 2) = \frac{P(A = 2 \mid E_i = .40)}{\sum_{i=1}^{21} P(A = 2 \mid E_i)} = \frac{.121}{1.092} = .111.$$

Decision

Table 2 gives the posterior probabilities that $\theta \leq .50$ for A = 2, 3, ..., 8 and the corresponding

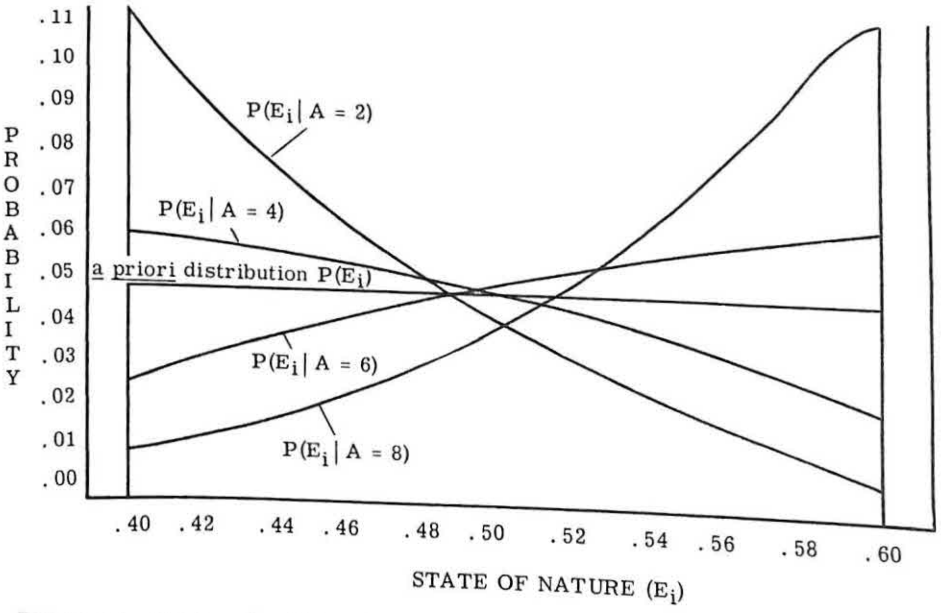


FIGURE 1. Rectangular Prior Probability Distribution, and Posterior Probability Distributions for A = 2, A = 4, A = 6, and A = 8.

TABLE 1

CONDITIONAL AND ACCUMULATED POSTERIOR PROBABILITIES

		A Priori Conditional Probabilities [$P(A = a_k \mid E_i)$]							Accumulated Posterior Probabilities [$\Sigma P(E_i \mid A = a_k)$]						
i	E_i	$P(2 \mid E_i)$*	$P(3 \mid E_i)$	$P(4 \mid E_i)$	$P(5 \mid E_i)$	$P(6 \mid E_i)$	$P(7 \mid E_i)$	$P(8 \mid E_i)$	$\Sigma P(E_i \mid 2)$	$\Sigma P(E_i \mid 3)$	$\Sigma P(E_i \mid 4)$	$\Sigma P(E_i \mid 5)$	$\Sigma P(E_i \mid 6)$	$\Sigma P(E_i \mid 7)$	$\Sigma P(E_i \mid 8)$
1	.40	.121	.215	.251	.201	.111	.042	.010	.111	.084	.061	.042	.027	.016	.009
2	.41	.111	.206	.250	.209	.121	.048	.012	.212	.165	.122	.085	.056	.038	.020
3	.42	.101	.197	.248	.216	.131	.054	.015	.305	.242	.182	.130	.088	.056	.034
4	.43	.093	.186	.246	.223	.140	.060	.017	.390	.315	.241	.177	.122	.080	.049
5	.44	.084	.177	.242	.229	.150	.067	.020	.467	.384	.300	.224	.158	.106	.068
6	.45	.077	.166	.238	.234	.160	.072	.023	.538	.449	.358	.273	.197	.134	.089
7	.46	.068	.156	.233	.239	.169	.082	.026	.600	.510	.414	.322	.238	.166	.113
8	.47	.062	.147	.227	.241	.179	.091	.030	.656	.568	.470	.372	.282	.202	.140
9	.48	.055	.137	.220	.244	.188	.099	.035	.707	.621	.523	.423	.327	.241	.172
10	.49	.049	.127	.213	.245	.197	.108	.039	.752	.671	.575	.474	.375	.283	.208
11	.50	.044	.117	.205	.246	.205	.117	.044	.792**	.717	.624	.525	.425	.329	.248
12	.51	.039	.108	.197	.245	.213	.127	.049	.828	.759	.672	.576	.477	.378	.293
13	.52	.035	.099	.188	.244	.220	.137	.055	.860	.798	.718	.627	.530	.432	.343
14	.53	.030	.091	.179	.241	.227	.147	.062	.887	.834	.761	.677	.585	.490	.400
15	.54	.026	.082	.169	.239	.233	.126	.068	.911	.866	.802	.727	.642	.551	.462
16	.55	.023	.072	.160	.234	.238	.166	.077	.932	.894	.841	.775	.699	.616	.533
17	.56	.020	.067	.150	.229	.242	.177	.084	.950	.920	.878	.823	.758	.685	.610
18	.57	.017	.060	.140	.223	.246	.186	.093	.966	.944	.912	.869	.818	.758	.695
19	.58	.015	.054	.131	.216	.248	.197	.101	.980	.965	.943	.914	.878	.835	.788
20	.59	.012	.048	.121	.209	.250	.206	.111	.991	.983	.973	.958	.939	.916	.889
21	.60	.010	.042	.111	.201	.251	.215	.121	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Σ	1.092	2.554	4.119	4.808	4.119	2.554	1.092							

*$P(2 \mid E_i) = P(A = 2 \mid E_i)$; .121 = probability of exactly two persons in the experimental group scoring above the mean of the traditional group if .40 of a population given "modern" mathematics would in fact score above the mean of the traditional group.

** .792 = $P(\theta \leq .50 \mid A = 2)$ = probability of at most .50 of a population given "modern" mathematics would score above the mean of the traditional group (H_0) if exactly two persons in the experimental group scored above the mean of the traditional group.

odds ratios. The odds ratio can be defined as the numerical chance that $\theta > .50$ after observing that $A = a_k$. Thus, the odds ratio is

$$\theta(a_k) = \frac{P(\theta_1 \mid a_k)}{P(\theta_o \mid a_k)} = \frac{1 - P(\theta_o \mid a_k)}{P(\theta_o \mid a_k)} .$$

For example, the odds ratio of $\theta > .50$ when $A = 2$ is

$$\theta(2) = \frac{1 - .792}{.792} = \frac{.208}{.792} = \frac{2}{8} = \frac{1}{4} .$$

At this point, the decision (D) can be made objectively.

D = (odds ratio) × (loss ratio).

If $D < 1$, the superintendent will decide that $\theta \leq .50$ and deny the request to adopt a "modern" mathematics program. If $D > 1$, he will decide that $\theta > .50$ and grant the request to adopt a "modern" mathematics program. If $D = 1$, he would reserve his decision until a few more cases were observed. Table 3 gives the D values for the various values of A. The D values in Table 3 indicate that if $A \geq 7$ the superintendent would decide in favor of the proposed "modern" mathematics program, but if $A < 7$ he would decide against it.

TABLE 2

POSTERIOR PROBABILITIES FOR θ CORRESPONDING TO VARIOUS VALUES OF A

$A = a_k$ [Observed Frequency]	$P(\theta \leq .50)$	$P(\theta > .50)$ $[1 - P(\theta \leq .50)]$	Odds Ratio $(\theta > .50)$
2	.792	.208	1:4
3	.717	.283	3:7
4	.624	.376	2:3
5	.525	.475	1:1
6	.425	.575	3:2
7	.329	.671	7:3
8	.248	.752	3:1

$P(\theta \leq .50) = P(\theta \leq .50 \mid A)$ = Accumulated Posterior Probabilities in $i = 11$ row in Table 1

TABLE 3

D VALUES FOR VARIOUS VALUES OF A

$A = a_k$	$L(\theta) \times \theta(a_k) = D$
2	(1/2) (1/4) = .125
3	(1/2) (3/7) = .214
4	(1/2) (2/3) = .333
5	(1/2) (1/1) = .500
6	(1/2) (3/2) = .750
7	(1/2) (7/3) = 1.170
8	(1/2) (3/1) = 1.500

DISCUSSION

In the illustrated problem, the Bayesian researcher would require $A \geq 7$ before rejecting H_o in favor of H_1. When using Bin (10, .48) a priori distribution with $\alpha = .05$, the classical researcher would require $A \geq 8$ before making the same decision. Thus, by not formally determining the loss function nor taking into consideration the posterior distribution and sticking close to the .05 level of significance, the classical researcher would be overly strict. In order to make the same decision as the Bayesian (who is taking these things into consideration) the classical researcher would have to set $\theta \geq .142$, an unlikely possiblity. In actuality, because of the seriousness of the action, the classical researcher might even require $\theta = .01$ and thus $A \geq 9$ for acceptance of H_1.

The Bayesian approach to statistical inference offers some definite advantages to the researcher over the classical approach. The former approach forces the researcher to determine the losses to be incurred if he makes a decision and is wrong. It allows the researcher to base his decision on the latest information (posterior distribution) while taking into consideration prior information without being restricted by a priori assumptions regarding the shape of the posterior distribution, such as the assumption of normality necessary for most classical designs. Furthermore, the final decision is objectively based on the product of the odds ratio and the loss ratio. Even though the Bayesian approach does have these advantages, it should not be considered a cure-all for statistical inference. In situations where the decision-maker is faced with choosing between two alternatives to a problem – one of which must be accepted – in which the decision will involve statistical hypothesis testing, the Bayesians have come up with a cogent technique. But in situations where the decision-maker is faced with choosing from among several alternatives, particularly when interacting variables might be involved, the Bayesians have not been convincing. Furthermore, they have not been satisfying in situations involving basic research where the purpose is to gain further information to integrate into a larger picture or in which there is not an action decision to be made.

REFERENCE

1. Marshall, Jon C., "Bayesian Decision," Journal of Counseling Psychology, XIV (July, 1967), pp. 342-345.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

COLLEGE STUDENT RATINGS OF INSTRUCTORS

NICHOLAS F. RAYDER
Michigan State University

ABSTRACT

In a great many of our institutions of higher learning, student and faculty groups are organizing systems for students to rate instructors. Often, the summarized results of these ratings are printed and disseminated throughout the campus. On several campuses, the instructor's ratings are incorporated into his personal file and may be used as a criterion in granting tenure or in awarding faculty promotions. Before student ratings of instructors are used in making decisions about a particular faculty member, several questions must be explored. Are student ratings related to such student characteristics as his grade point average (GPA) or his class level? And, what instructor characteristics are associated with high ratings? In a study to examine the nature of student ratings over five thousand college students rated their instructors on each of three criteria. These ratings were then correlated with selected student characteristics and instructor characteristics. The technique used was multiple linear regression. Generally, mean student ratings of instructors were more related to instructor characteristics than to such student characteristics as age, grade level, GPA , or course grades.

FRED M. HECHINGER, education editor of the New York Times has written:

> Students have persuaded the educational leadership that the student's voice on campus matters deserves to be heard and that their complaints about inadequate teaching and a curriculum out of step with contemporary society are often justified (2).

He further states that students now seek to sit on educational policy-making committees, a responsibility that faculty members often shirk.

These student complaints are dramatized by unrest on campuses throughout the country. Students at large universities complain about being disturbed and somewhat threatened by the dehumanization and the loss of identity they perceive in the multiversity. Many are deeply concerned about the nature of our system of higher education and want change, most noteworthy in the improvement of instruction.

Frederick Rudolf, professor of history at Williams College, finds nothing new in the desire of students to bring about change and reform on American campuses. He states: "For a few years the college is the oyster and they (students) will have it served up exactly as they wish it, unless there are those who help them to some other, perhaps wiser choice (2). Some faculty and administrators are helping by cooperating with student organizations wanting to improve instruction. At the City College of New York, a joint student-faculty committee was established to develop in a truly collaborative fashion some type of course evaluation. A program evolved consisting of the distribution of course-instructor questionnaires to instructors on a voluntary basis. The instructor distributed the scales, collected them, and was free to use them in any way he wished. He could send them to the office of the Chairman of the Department of Student Services where a confidential report would be prepared and returned to him. Of 400 instructors requesting to participate, only 144 sent the collected data to the Department of Student Services for scoring.

Other faculty groups are responding to students who want to have a greater role in the affairs of their institution. In 1965, Cornell University appointed a Committee on the Quality of Undergraduate Instruction chaired by Dr. Alfred E. Kahn (3). Although appointed by the executive branch of the university, the members worked closely with the student body to study the academic aspects of undergraduate education at Cornell. The comprehensive report that resulted from the study included the following findings:

1. Grossly inadequate teaching occurs in more instances than is tolerable.
2. There can be no doubt that student dissatisfaction with undergraduate instruction has basis in fact.

3. There is only one ultimate determinant of the quality of undergraduate instruction and only one ultimate source of its improvement—the individual teacher himself.

The committee recommended unanimously that "teachers obtain anonymous student evaluations of their courses" and, with one member dissenting, "that the university give Student Government financial and technical assistance in preparing and publishing competent and responsible campus-wide student evaluations of courses" (3).

A similar committee at Michigan State University, The Committee on Undergraduate Education, was appointed in 1967 by the school's president to reexamine the undergraduate program and to propose desirable changes. In a section of their final report concerning student evaluation of teaching, the committee stated that "an instructor should know when students do not approve of a particular technique or approach (1). The committee recommended that a special all-university office be given the responsibility and adequate support: to revise or develop new Student Instructional Rating Forms; to collect and disseminate normative data relative to these forms; and to analyze the results and prepare summary reports." They further recommended that "Faculty members be encouraged to have these summary reports sent to their department chairman so that the results can be involved in decisions pertaining to them."

The number of colleges and universities using some method of student evaluation is increasing rapidly and not without some adverse consequences. Obviously, there are many dangers when student rating of faculty performance may be a factor in a teacher's professional standing, possibly affecting his promotion.

If studies are not methodologically sound, popularity might rate higher than intellectual rigor; or political views might influence decisions that ought to be arrived at by a process of free inquiry.

Other questions concerning the validity of student ratings are equally critical, and this report summarizes a study designed to examine several student and instructor characteristics to see what relationship each had with student ratings of their instructors.

The research was carried out within the School of Education at Colorado State College (Greeley, Colorado) during the 1967 winter quarter. The specific instructor population was identified as all faculty within the School of Education who were teaching courses during winter quarter. The student population was all those who were enrolled in those courses.

PROBLEM

No attempt was made to identify effective and ineffective instructors. Instead, the study concentrated upon variables believed to have some relationship to student ratings. Therefore, the major questions studied were:

1. Does student sex, age, major, level of education, GPA, or course grades previously received from the instructor being rated have any relationship to ratings of instructors?

2. Are instructors who differ in sex, age, faculty rank, degree held, major area, or length of teaching experience rated differently by students?

A further question concerned the accuracy of students' reports of their GPA's. This was done to check the validity of student information. If large discrepancies existed between students' reported GPA and their actual GPA (determined from college records) the remaining information provided by students would be suspect.

METHODOLOGY

To collect the necessary data, a 1-page Student Scale was constructed. The scale asked students to provide information about themselves and to rate their instructor on each of three 7-point continuums. The referents on each continuum were identified by Ryans as being important teacher traits or patterns of behavior (4). These were:

I. aloof, egocentric, restricted behavior	vs.	friendly, understanding, behavior
II. evading, unplanned, slipshod behavior	vs.	responsible, systematic, businesslike behavior
III. dull, routine behavior	vs.	stimulating, imaginative enthusiastic behavior

A short Instructor Characteristics Scale was designed to obtain from faculty such information as sex, age, rank, degree held, major area, and length of teaching experience.

Next, a group of doctoral students were trained to administer the scales. At the instructors' convenience, 4,402 Student Scales were administered in the existing classes taught by eighty-seven different instructors. At the same time, the Instructor Characteristic Scale was administered.

Correlation analyses were performed to answer the questions. One statistical technique used to analyze the data was multiple regression as presented by Ward and Bottenberg (6). In this approach the contribution of a variable or a set of variables to estimation of a criterion was measured by the difference between two multiple correlation coefficients (RSQ's); one obtained for a regression model in which all the predictors were used, and the other (RSQ) obtained for a regression model in which the variable or subset of variables under examination was deleted (5).

FINDINGS

Based on statistical analysis of data collected on eighty-seven instructors from 4,285 usable Student Scales, the following conclusions were reached:

TABLE 1

CORRELATIONS AND RSQ's BETWEEN CRITERIA AND STUDENT CHARACTERISTICS

	Student Characteristic (s)					
Rating Continuum	Reported GPA		Previous Grade (s)		(Sex, Age, Level, Major)	
	N	r	N	r	N	RSQ
I	4,032	-.03	753	.14	4,285	.006
II	4,032	-.02	753	-.01	4,285	.015
III	4,032	.03	753	.18	4,285	.019

1. Students remembered and accurately reported their GPA's. The correlation for one hundred randomly selected student-reported GPA's and their actual GPA's was .96.

2. Student ratings of instructors were not substantially related to student's sex, age, grade level, major area, GPA or grade(s) previously received from the instructor they rated (Table 1).

As indicated in Table 1, the highest correlation (.18) was obtained between students previous grade(s) earned from the instructor they were rating and their rating of that same instructor on continuum III. Also, there was no relationship, reflected by correlations of -.03, -.02, and .03 between students GPA and their ratings.

When the student characteristics of sex, age, grade level, and major were used collectively to predict ratings, the RSQ's were all less than .02. To demonstrate the low predictability of this set of characteristics, the regression weights associated with each characteristic were used to predict how two hypothetical students would have rated an instructor. These predictions are found in Table 2.

Based on the regression equation, in which sex, age, level of education, and major area were used, Student No. 1, a 35-year-old male graduate student majoring in educational adminstration would be predicted to rate an instructor 3.5 on continuum I, 2.7 on continuum II, and 2.7 on continuum III. Another student (No. 2), an 18-year-old female majoring in elementary education was predicted to rate an instructor 3.3, 2.8, and 2.3 on the three respective continuums. These small differences in predicted ratings (0.2, 0.1, and 0.4) for quite dissimilar students demonstrated the low relationship between this group of student characteristics and student ratings of instructors.

TABLE 2

PREDICTED INSTRUCTOR RATINGS FROM STUDENT CHARACTERISTICS FOR TWO HYPOTHETICAL UNLIKE STUDENTS

Characteristics of Student					Predicted Rating on Continuum		
Student	Sex	Age	Level of Education	Major	I	II	III
1	Male	35	Graduate	Educ. Admin.	3.5	2.7	2.7
2	Female	18	Freshman	Elem. Educ.	3.3	2.8	2.3
				Difference	0.2	0.1	0.4

TABLE 3

RSQ's FOR INSTRUCTOR CHARACTERISTICS AS PREDICTORS OF STUDENT MEAN RATINGS ON CRITERIA FOR EIGHTY-SEVEN INSTRUCTORS

Model	Instructor Variables	Variables Deleted	Continuum Predicted	RSQ Value
1	Sex, Age, Rank, Degree, Dept. Years Exp.	None	I	.15
			II	.27
			III	.18
2	Age, Rank, Degree, Dept., Years Experience	Sex	I	.15
			II	.27
			III	.18
3	Sex, Rank, Degree, Dept., Years Experience	Age	I	.15
			II	.19
			III	.18
4	Sex, Age, Degree, Dept., Years Experience	Rank	I	.15
			II	.24
			III	.18
5	Sex, Age, Rank, Dept., Years Experience	Degree	I	.15
			II	.27
			III	.18
6	Sex, Age, Rank, Degree, Years Experience	Dept.	I	.05
			II	.14
			III	.08
7	Sex, Age, Rank, Degree Department	Years Teaching Experience	I	.14
			II	.25
			III	.16
8	Sex, Dept.	Age, Rank, Degree, Years Experience	I	.11
			II	.13
			III	.13

3. Generally, mean student ratings of instructors showed more relationship to instructor characteristics (Table 3).

   For the instructor characteristics of sex, age, degree held, faculty rank, and number of years experience, collectively, the RSQ was .27 for continuum II ("evading" vs. businesslike"). This means that 27 percent of the variability of ratings on continuum II is related to this set of instructor characteristics.

4. The greatest reduction in the index of predictability (RSQ) was observed when the "department" classification variable was deleted from the equation (Model 6). The RSQ value for continuum I dropped from .15 to .05; for continuum II from .27 to .14; and for continuum III from .18 to .08.

   Mean student ratings according to major areas reflected this reduction in RSQ's. For example, means calculated for each department on the "aloof vs. friendly" continuum ranged from 4.0 to 5.8, on the "unplanned vs. systematic" scale from 3.9 to 5.4, and on the "dull vs. stimulating" scale from 4.1 to 5.2.

TABLE 4

INTERCORRELATION COEFFICIENTS BETWEEN INSTRUCTOR CHARACTERISTICS AND THEIR AVERAGE RATING BY STUDENTS

Variable	1	2	3	4	5	6
(1) Continuum I						
(2) Continuum II	.47					
(3) Continuum III	.71	.67				
(4) Age	-.17	-.15	-.25			
(5) Faculty Rank	-.15	.15	-.06	.51		
(6) Highest Degree Held	-.16	.12	-.05	.37	.79	
(7) Years Teaching Experience	-.18	.00	-.27	.86	.54	.37

5. The correlations of the three student rating criteria with instructor age, faculty rank, highest degree held, and years of teaching experience yielded twelve correlations varying from -.27 to +.15 (Table 4).

   Nine of these were negative, suggesting that older, more educated and more experienced instructors were rated lower than younger and less experienced instructors. Further, because *mean* student ratings were used as an indicator of each student's rating, correlations were attenuated. If each student's *actual* rating scores were used, correlations would have been higher and the inverse relationship magnified.

DISCUSSION

To involve students in any system of faculty evaluation seems perilous. Yet it is students who have the most contact with instructors. To ignore them seems even more perilous. The results of this study confirmed that student ratings are not influenced even by variables such as grades previously earned from the instructor being rated. Since student ratings were found to be relatively unbiased, more confidence could be placed on the results of how certain instructor characteristics were reflected in students' ratings. As was mentioned, these results suggested that ratings were different from various departments within the School of Education.

Further, the younger instructors with less faculty rank, and fewer years of teaching experience were consistently rated more "friendly, understanding; systematic, responsible, businesslike; stimulating, imaginative, and enthusiastic."

Probably as important as the specific findings, the investigator made several observations that did not lend to statistical analysis:

a. The multiple regression approach used in this study was effective in analyzing the degree to which certain student and instructor characteristics were reflected in student ratings.

b. Ryans (4), using factor analysis, identified three patterns of teacher behavior in the *Characteristics of Teachers* study. However, in the present study, when each of these patterns was placed on a rating continuum and administered to college students, the intercorrelations among ratings were high (Table 4). This suggested that the traits Ryans identified as being independent do overlap. Therefore, a total student rating score of an instructor could have been obtained by adding the student ratings made on each continuum.

c. The study gave a large portion of the student body a chance to rate their instructors. Often, when such a route is not provided or such a desire is thwarted, the upshot may be a poorly organized student-directed evaluation resulting in a faculty "blurb" sheet.

d. After student leaders in charge of developing an instructor evaluation program talked with this investigator, they appreciated the time, sophistication, and financing needed to carry out a methodologically sound student evaluation. And, after examining the complexity of the issues, these student leaders agreed that some degree of confidentiality should be maintained and that student publications on course evaluations should contain more objective information such as the number and type of exams given.

SUMMARY

As a consequence of the study reported here, we now have more information on the nature of college students ratings of their instructors, at least at Colorado State College. We have shown that some information reported by students is accurate. We have shown that student ratings are not influenced by such variables as previous grades received from the instructor they rated. And we have shown that instructors who differ on certain characteristics are rated differently by their students.

Still other equally important issues remain unanswered. We need to know what student characteristics, if any, do influence their rating of instructors. We need to know if a parallel study performed within another academic area or at another institution will yield similiar results. And, we need to know if there are other instructor characteristics that better reflect student ratings.

Yet, in terms of students rating their instructors the question: "to rate or not to rate" can best be answered by citing a quotation attributed to Aristotle: "You get a better notion of the merits of the dinner from the dinner guests than you do from the cook."

REFERENCES

1. Adams, Arthur E., "Improving Undergraduate Education: Report of the Committee on Undergraduate Education" (Michigan State University, East Lansing, Michigan, 1967).
2. Hechinger, Fred M., "Education: Now Grades for Teacher," The New York Times Sunday, October 24, 1965, Section E, page 9.
3. Kahn, A. E., (Chairman) Report to the Faculty Committee on the Quality of Undergraduate Instruction (unpublished report), Cornell University, Ithaca, New York, 1965.
4. Ryans, D. G. Characteristics of Teachers, (Washington, D. C.: American Council on Education, 1960).
5. Ward, J. H. Jr., Multiple Linear Regression Models, Computer Applications in the Behavioral Sciences, Harold Borke (ed.), (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1962).
6. Ward, J. H. Jr., and Bottenberg, R. A., Applied Multiple Linear Regression, Technical Report PRL-TDR-63-6, 6570th Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Lackland Air Force Base, Texas, 1963.
7. Werdell, Philip, Course and Teacher Evaluation (Washington, D. C.: United States National Student Association, 2115 S. Street N. W. Zip 20008, 1967).

ders

DEMBAR
EDUCATIONAL
RESEARCH
SERVICES, INC.

POST OFFICE BOX 1148 • MADISON WISCONSIN 53701

PERTINENT RECENT TITLES FROM DERS

THEATER IN AMERICA

By Prof. Robert E. Gard, Prof. Marston Balch, and Pauline Temkin
212 pages $6.50 hardcover, $4.95 softcover
The most complete story of contemporary theater published to date. Commissioned by the National Theater Conference.

FUN WITH GAMES

By Prof. John E. Anderson
188 pages $4.95
If a suitable recreational activity isn't described and explained in this book, it probably doesn't exist.

WISCONSIN SIDEROADS TO SOMEWHERE

By Prof. Clay Schoenfeld
266 pages $2.95
A collection of stories and essays in the spirit of Thoreau, Muir, and Leopold. Offers entertaining, informative reading about the adventures in outdoor recreation and conservation that are everybody's for the finding down the sideroads of America.

RESEARCH AND DEVELOPMENT TOWARD THE IMPROVEMENT OF EDUCATION

Edited by Profs. Herbert J. Klausmeier and George T. O'Hearn
176 pages $7.25 hardcover, $5.75 softcover
A compilation of papers by 21 leading educators devoted to imaginative strategies for cognitive-learning research and the resulting instructional practices.

IDEAS AND IMAGES

By Prof. Lindley J. Stiles
96 pages $3.50
Popular poetry about and for young people—and for all who share with youth the exciting venture of growing up.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

PREDICTION OF COMMITMENT TO THE CHEMICAL BOND APPROACH TO TEACHING

CYNTHIA V. L. WARD[1]
Rhode Island Junior College

ABSTRACT

In general terms, the problem under investigation is that of determining whether selected teacher characteristics can be used as predictors for the commitment of teachers to a certain approach to the teaching of high school chemistry (Chemical Bond Approach). Teachers attending a summer institute were initially surveyed to determine these characteristics; included in this group were expectations of role related groups, innovativeness characteristics, and personal characteristics. At the completion of the institute, the participants were tested regarding their attitudes toward and their willingness to use the approach. Employing the stepwise regression technique of analysis, two characteristics (perceived disposition of reference group and innovativeness) were found as strong predictors. Pre-institute characteristics were also shown to account for a significant portion of the commitment of teachers to the approach.

THE PROBLEM

Since the inception in 1953 of National Science Foundation summer institutes, hundreds of thousands of high school teachers have been exposed to new developments in the teaching of science and mathematics under the auspices of these programs. Despite the size of this effort it would be difficult to gauge the impact that these summer institutes have had on the evolution of high school science teaching in the United States. The determining factor controlling the change attributable to these institutes is the decision made by the individual participant on his return from an institute as to the modifications he may make in his own teaching in light of the summer's experience.

It was the purpose of this study to examine a group of teachers faced with a contemplated change in their teaching approach as a result of participating in a common introduction to the teaching of high school chemistry by the Chemical Bond Approach (CBA). Attention was focused on how the teachers reacted in this choice-situation, and whether certain characteristics of the teachers prior to their involvement in the institute could be used to predict the choice that they would make.

The mode of investigation used in this study was essentially empirical in nature. Using one group of institute participants, a measure of pre-institute characteristics was determined, as well as a measure of the extent to which these teachers appeared to be influenced by the offerings of the institute. Research from related fields provided indications as to the most promising pre-institute characteristics to be considered. The study was conceptualized within the framework of decision-making theory (5, 6), and associated research results on the adoption of an innovation (7, 9), planned change (1, 4), role theory (2), and reference group theory (8) were used to further construct the theoretical basis.

METHOD

The decision makers in this experimental setting were the high school chemistry teachers attending a CBA institute sponsored by the National Science Foundation in the Summer of 1965. The pre-institute characteristics of the teachers were considered to be possible determinants in affecting the decision in this situation. Various degrees of commitment on the part of the teachers to the CBA method of teaching high school chemistry at the completion of the institute were selected as the alternatives of the decision process (see Figure 1).

Due to the fact that the CBA institute provided only the vehicle for the decision, it was not necessary to explore the course offering in detail. It is sufficient to say that the indications were that teachers completing the course in the past had mixed feelings as to the applications of this approach in their own teaching, thus the possible alternatives of this decision situation were expected to vary along a continuum ranging from total rejection to total commitment to the CBA method. This research project was not designed to include evaluations of the CBA institute attended by the sample or the CBA method of teaching high school chemistry. Ergo, any inferences concerning such evaluations based on the data collected in this study are unwarranted.

Determinants which were investigated as possible predictors for the choice-situation were divisible into three distinct sets: A. the expectations of the groups in role relationship to the high school chemistry teacher; B. the innovativeness of the teacher and his teaching environment; and C. personal characteristics of the teacher. The listing below presents the selected determinants hypothesized to be indices of commitment to the CBA method. In hypotheses testing these factors served as independent variables.

A. Perceived Expectations of Role Related Groups

   1. Membership group
   2. Other direct contact group
   3. Reference group
   4. Authority group
   5. Composite group
B. Factors Related to Innovativeness
   1. Innovativeness of school administration
   2. Pre-institute use and attitude toward CBA materials
   3. Pre-institute use and attitude toward CHEM Study materials
   4. Degree of professional activity (reading and attendance at meetings)
C. Personal Characteristics
   1. Sex
   2. Age
   3. Credit hours earned in chemistry
   4. Credit hours earned in physical chemistry
   5. State in which teacher was employed

The criterion or dependent variable, considered to be commitment to CBA, was conceptualized as being composed of an attitudinal component and a behavioral readiness component. To assess the attitude toward the CBA method of teaching high school chemistry, the individual teachers were asked to respond to a series of evaluative statements concerning the method. Behavioral readiness was gauged by the response made by the teacher when presented with a simulated situation in which a choice concerning the use of the CBA method was necessitated. A high degree of correlation was found, as expected, between the attitude and behavioral readiness components of the commitment variable.

PROCEDURES

The Sample

The sample was composed of 45 chemistry teachers from American public and independent schools. The sample was not assumed to be representative of the universe of high school chemistry teachers, or of high school chemistry teachers attending CBA institutes. Consequently, care should be exercised in generalizing from the results, though the results should provide indicators as to the reactions of other groups in similar choice-situations.

Development of the Research Instruments

Two written questionnaires were selected as the most effective and efficient means of collecting the data for this project. Questionnaire I contained items seeking to determine: 1) the background variables used to describe the sample, and 2) the pre-institute characteristics of the participants. Questionnaire II contained items aimed at distinguishing the extent of the teacher's commitment to the CBA method at the completion of the institute.

Both questionnaires were subjected to a pre-test conducted by correspondence, employing the 1964 participants of the same CBA institute. The results of the pre-test were used to revise the questions in an effort to reduce ambiguities and possible misinterpretations.

Questionnaire I: As indices to the first group of independent variables, the teachers were asked to respond to a series of items which sought to discover their perceptions of the dispositions of others to the CBA course. Five major groups (high school chemistry teachers, high school science supervisors, high school principals, college chemistry teachers, and science educators) were selected as groups which the teachers would anticipate to hold expectations for the use of CBA materials in the high schools. A series of items was included in the first questionnaire to determine which of these groups served as reference groups (groups to which the individual aspires to attain or retain membership)[2] and authority groups (groups which are recognized by the individual as having special knowledge in the area) for the individual teacher. Using these items, the dispositions toward CBA perceived by the teachers for their membership group, other direct contact group, reference group, authority groups, and the composite groups were determined. Other items were included on this questionnaire to measure the innovativeness factors and the selected personal characteristics of the teacher. Some sample questions from Questionnaire I are presented in Example 1. A set of arbitrary scales was designed to assign scores to the subjects on each of the independent variables.

Questionnaire II: The set of items devised to measure commitment to CBA was differentiated for three general types of high school chemistry students (college-bound science, college-bound nonscience, and terminal high school), because of the anticipated difficulty of treating high school chemistry students as a homogeneous group. With the use of the two components (attitude and behavioral readiness) and the three groups, six scales were obtained to represent the commitment variable. In order to retain information, these scales were not reduced to one, though they were highly correlated.

In the final analyses 13 items were used to measure the attitudinal component[3] and eight items were used for the behavioral readiness component. See Example 2 for samples of these items.

The procedure of principal-component analysis was used to establish that the items for each of the two components for the three different groups of students contributed to unidimensional scales. For the attitudinal component one factor accounted for at least 42 percent of the total variance under analysis; for the behavioral readiness component a minimum of 58 percent of the variance could also be attributed to this same factor. (See Table 1.) Because this factor accounted for close to one half of the variance for each of the three groups for both components, and because the other individual factors accounted for only a small part of the variance, only this major factor was used in subsequent analyses. It was assumed that the major factor was synonymous with favorableness toward CBA, and that this scale was, in fact, unidimensional.

A T score technique was utilized to assign each of the subjects a score on each of the six scales of the dependent variable. The T score is a special case of the standard score which employs a mean of 50 and adopts as its unit, one-tenth of a standard deviation. (3)

EXAMPLE 1

SELECTED ITEMS FROM QUESTIONNAIRE I

Assume that you are qualified for all of the following positions. Which one would be your first choice for a position at the present time?

a) secondary school chemistry teacher
b) secondary school science supervisor
c) secondary school administrator
d) college teacher of science education
e) college chemistry teacher

From which one group would you most value a compliment or a word of praise on your teaching of high school chemistry?

a) associate chemistry teachers
b) science supervisor
c) high school principal
d) science educators
e) college chemistry teachers

From which one of the following groups would you most likely seek advice in dealing with a curriculum problem for your high school chemistry class?

a) associate chemistry teachers
b) science supervisor
c) high school principal
d) science educators
e) college chemistry teachers

What opinions do you think that most high school chemistry teachers hold for the use of CBA materials in the high school?

a) favorable
b) neutral
c) unfavorable
d) no impressions

Which of the following statements best expresses the reaction of the principal and assistant principals in the school in which you teach toward developments in curriculum?

a) Desire the school to be at the forefront in using and testing experimental curriculum.
b) Are willing to accept curriculum developments if they have proved successful in trials in other schools.
c) Are receptive to developments that are already in use in the majority of schools.
d) Are very reluctant to make changes in curriculum.

What use, if any, have you already made of the CBA materials in your own teaching?

a) none
b) slight use
c) moderate use
d) extensive use
e) taught entire course

Indicate the extent to which you are familiar with the CBA test.

a) have read the text
b) have scanned the text
c) have not seen the text

ANALYSIS

A stepwise regression technique was used to determine which of the independent variables were significant predictors for commitment to CBA. Each of the six scales of the dependent variable was used individually with the set of independent variables for six separate analyses, and comparisons were made of the six sets of results.

A cut-off value for the $\underline{F}$ ratio was selected so that only those independent variables whose beta weights were significant, as determined by use of a $\underline{t}$ test with a .10 level of significance, were included in the regression equation. The data obtained from the analyses appear in tabular form in Table 2.

RESULTS

Two of the pre-institute factors emerged from the analyses as strong predictors on the majority of the six scales of commitment; these factors were the perceived disposition of the teacher's reference group and innovativeness. Each of these factors was in evidence in all but the behavioral readiness component – college-bound science student group, where the reference group disposition factor did not appear.

The innovativeness factor was redefined after the analyses as a composition of three of the sub-groups originally conceived as contributing to this factor. These variables (including: 1. the perceived innovativeness of the home administration in curriculum matters; 2. prior use of CBA materials; and 3. prior use of CHEM Study materials) were found to exhibit a duplication effect because of the high positive correlation between them. Thus, the removal of one of these variables from the correlation matrix resulted in a lowering of the partial correlation of those variables in the group remaining in the matrix. Consequently, these variables were considered as a group and were labeled for future reference as the innovativeness variable.

Other independent variables which were significant predictors for one or two of the six measures of commitment were: 1. whether the teacher was employed in New York State; 2. the number of undergraduate and graduate credit hours in chemistry earned by the teacher; and 3. the amount of professional reading and number of professional meetings attended by the teacher.

Teaching in the State of New York was a variable which was represented twice on the attitudinal component as a predictor. If a teacher was employed in New York he was less likely to be strongly committed to the CBA method than were teachers from other states. Comments by institute participants reinforced the suspicion that teachers from New York felt restricted in the extent to which they could improvise upon the syllabus set for the New York Regents Examination. These same individuals expressed the hope that the revised syllabus, which was scheduled for introduction in 1968, will include certain topics found in the CBA presentation.

The last two variables proved to be negative predictors; thus they were significant predictors in the opposite direction than that originally hypothesized. The greater the number of hours of chemistry studied by the teacher, the less likely the teacher was to be committed to CBA. This reverse-predictor operated only for the college-bound science and non-science groups for the behavioral readiness measure. A possible rationalization for these results is that teachers who were already familiar with the concepts introduced in the CBA course had included these concepts in their own teaching prior to their institute experience. Therefore, they felt it unnecessary to adopt the CBA course to supplant a course which already included much of the same material. Those teachers with weaker backgrounds in chemistry, meeting certain concepts in the CBA course for the first time, might be more prone to adopt the new method in its entirety. For the behavioral readiness – college-bound science group, the second negative predictor, professional reading and meetings attended emerged as significant. From the data collected in the course of this study, no logical explanation could be advanced as to why this predictor should have been operative in the reverse direction to that originally hypothesized.

Table 3 indicates the importance of the best predictors in explaining the variance in the commitment variables. Between one-fourth to two-fifths of the total variance is attributable to the pre-institute characteristics which are strong predictors.

CONCLUSIONS

A primary conclusion to emerge from this study was that the perceived disposition of the teacher's reference group toward a teaching method is a salient predictor to the teacher's commitment to that method. This result is consistent with previous research in related areas, which lends credibility to the

EXAMPLE 2

SELECTED ITEMS FROM QUESTIONNAIRE II

Attitudinal Component

(Teachers were to respond to each item for each of the three groups of high school students using a five-point "agreement" scale, ranging from strongly agree to strongly disagree.)

The CBA course gives the students sufficient grounding in the descriptive side of chemistry.

The CBA course provides a stimulating experience for the majority of the students.

The CBA course is so designed that the students can relate previous and new learnings effectively and use this information in dealing with new problems.

A laboratory, such as that presented in CBA, which requires the students to perform the functions characteristic of a chemist is inappropriate for the students.

Behavioral Readiness Component

(A simulated situation was devised detailing a hypothetical teaching situation in which the three different groups of high school students were represented. The teachers were asked to respond to the following questions in terms of what they would do if they were confronted with these decisions in the presented setting.)

Approximately what percentage of the teaching time in the classroom situation would you use CBA material?

Indicate what portion of the laboratory experiments you would have the students do from each of the group divisions in the CBA laboratory book.

Would you recommend that the school system purchase sufficient copies of the CBA so that it could be used as a textbook?

TABLE 1

RESULTS OF THE PRINCIPAL COMPONENT ANALYSES

	College-Bound Science Group Factors			College-Bound Non-Science Group Factors			Terminal High School Group Factors			
	1	2	3	1	2	3	1	2	3	4
Attitudinal Component										
Eigen-value	7.8	1.6	1.2	8.2	1.6	1.3	6.9	1.7	1.3	1.2
Percent Trace	48.8	10.3	7.5	48.1	9.3	7.4	42.8	10.4	7.9	7.3
Behavioral Readiness Component										
Eigen-value	5.4			5.1			4.7			
Percent Trace	67.7			63.8			58.7			

TABLE 2

INDEPENDENT VARIABLES SELECTED AS BEST PREDICTORS FOR THE SIX SCALES OF THE DEPENDENT VARIABLE

College-Bound Science Group	Beta Weights*	College-Bound Non-Science Group	Beta Weights	Terminal High School Group	Beta Weights
Attitudinal Measure					
Innovativeness	.40	Reference Group Disposition	.45	Reference Group Disposition	.41
Reference Group Disposition	.30	Innovativeness	.34	Innovativeness	.40
Teach in New York State	-.24	Innovativeness	.21	Teach in New York State	-.26
Behavioral Readiness Measure					
Chemistry Hours	-.34	Reference Group Disposition	.33	Innovativeness	.38
Innovativeness	.31	Innovativeness	.33	Reference Group Disposition	.35
Readings and Meetings	-.30	Chemistry Hours	-.30		

*Only predictors with significant beta weights ($p < .10$) are reported

TABLE 3

VALUES FOR THE SQUARE OF THE MULTIPLE CORRELATIONS (R^2)* OF THE BEST PREDICTORS FOR THE SIX SCALES OF THE DEPENDENT VARIABLE

Scale	R^2
Attitudinal Component College-Bound Science Group	.31
Attitudinal Component College-Bound Non-Science Group	.41
Attitudinal Component Terminal High School Group	.39
Behavioral Readiness Component College-Bound Science Group	.23
Behavioral Readiness Component College-Bound Non-Science Group	.33
Behavioral Readiness Component Terminal High School Group	.24

*R^2 states that proportion of the total variance which can be attributed to the independent variables.

present study. Sherif (10) stated that "...numerous studies coming both from psychologists and from sociologists have shown that the major source of the individual's weighty attitudes are the values or norms of the groups to which he relates himself," (i. e. his reference groups). The expectations of other groups (membership group, direct contact group, authority group, and composite groups) within the role pattern of the teacher were shown in this study not to be significant predictors for the commitment of the teacher to the CBA method.

The second strong predictor indicated in the analyses was an innovativeness variable. Teachers ranking high on the innovativeness variable were more likely to evidence strong commitment to the CBA method, than were teachers with a lower innovativeness score. This factor was conceived as an indicator of the extent to which the individual teacher and his immediate school environment demonstrated the practice of adopting new ideas within the professional sphere. Little attempt was made in measuring this variable to separate the innovativeness of the teacher from the pressures toward innovativeness exerted within the school. However, since this variable was demonstrated to be an efficient predictor, the desirability of exploring the secondary question is enhanced.

Other predictors were also shown to be significant for some measures of the commitment variable, but did not exhibit the universality of the two primary predictors. Predictors in this group included whether the teacher was employed in New York State, the teacher's number of earned hours of credit in chemistry, and the degree to which the teacher attended professional meetings and read professional journals. These last two predictors exhibited a reverse relationship to that which had been originally hypothesized.

The results of this investigation also emphasize that these pre-institute characteristics of the teachers accounted for a significant portion of the variance in the teachers' commitment to CBA. Because of the importance of these characteristics in the final decision, those sponsoring and directing summer institutes might consider modifications in the institute offerings for different types of participants.

FOOTNOTES

1. This manuscript is based on the writer's doctoral dissertation completed under the direction of Dr. Fletcher G. Watson and Dr. Maurice Belanger, "A Study of Initial Characteristics of Participants in a CBA Summer Institute as Predictors for Final Commitment to CBA," Graduate School of Education, Harvard University, 1966.

2. Eight items were devised to ascertain the commitment of the teacher to a particular group, the identification groups of the teacher, the aspirations of the teacher to group membership, and the group most respected by the teacher. Some of these items were adapted from the work of Weller (11).

3. Because of the possibility that students in the college-bound non-science group might be either continuing or terminal chemistry students, an additional item was included in the attitudinal component for this group, thus raising the total to 14 items.

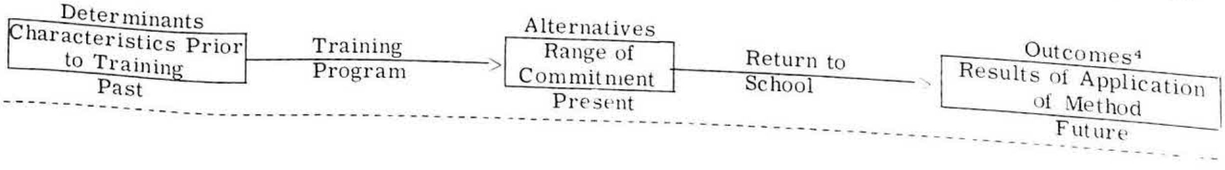


FIGURE 1

SCHEMA REPRESENTING THE DECISION-MAKING PROCESS OF TEACHERS EXPOSED TO AN INSTITUTE TRAINING PROGRAM

4. As an extension of the major portion of the study, a follow-up questionnaire was sent to the participants after they returned to their schools. The purpose of this questionnaire was to determine the degree to which these teachers utilized the CBA method in their own teaching. Failure to receive a significant percentage of returns made this analysis unfeasible.

REFERENCES

1. Bennis, Warren G., Beene, Kenneth D., and Chin, Robert (Editors) The Planning of Change (New York: Holt, Rinehart and Winston, 1961).
2. Gross, Neal, Mason, Ward S., and McEachern, Alexander W., Explorations in Role Analysis (New York: John Wiley and Sons, 1958).
3. Guilford, J. F., Fundamental Statistics in Psychology and Education (New York: McGraw-Hill, 1965), p. 523.
4. Lippitt, Ronald, Watson, Jeanne, and Westley, Bruce, The Dynamics of Planned Change (New York: Harcourt, Brace, and World, 1958).
5. March, James G., and Simon, Herbert A., Organizations (New York: John Wiley and Sons, 1958).
6. Meeker, Robert J., Shure, Gerald E., and Miles, S. Rogers, "Research Approach to Complex Decision Making," Educational Research: New Perspectives (Danville, Ill.: The Interstate Printers and Publishers, 1963), pp. 165-187.
7. Miles, Matthew B. (Editor) Innovation in Education (New York: Bureau of Publications, Teachers College, Columbia University, 1964).
8. Newcomb, Theodore M., Social Psychology (New York: Dryden Press, 1950).
9. Rogers, Everett M., Diffusion of Innovations (New York: Free Press of Glencoe, 1962).
10. Sherif, Muzafer, "The Concept of Reference Groups in Human Relations," Group Relations at the Crossroads (New York: Harper and Brothers, 1953), p. 207.
11. Weller, Charles M., "The Role Orientation of the Secondary School Science Teacher," unpublished Ed.D. thesis, Graduate School of Education, Harvard University, 1965.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

THE USE OF EXPERIMENTAL DESIGNS IN THE DECISION-MAKING FEEDBACK PROCESS

GORDON WELTY
The Board of Public Education, Pittsburgh, Pennsylvania

ABSTRACT

This research paper is concerned with the relationship of feedback of research findings and with educational decision making.

STUFFLEBEAM has argued that on methodological grounds, experimental designs are incompatible with feedback of research findings for educational decision making. A change in treatment during the evaluation cycle vitiates the research findings sufficiently to render rigorous controls useless (1). If it is true that there is a logical incompatibility, then either feedback for program change in the social action projects is impossible, or else experimental designs must be abandoned and a new "theory of evaluation" developed. We show that his contention is ill-founded. Perhaps on practical grounds, a design is incompetently implemented, or resources are lacking, etc. However, there is no basis in principle for not using experimental designs, and having feedback during the evaluation cycle.

Represent a project to be evaluated as

$$y_o = \Phi(x)$$

when x is input, Φ is treatment or process, and y_o is a time independent output. The experimental design (e.g. Campbell and Stanley's Design 4) is symbolized by

$$y = \Phi[t(x)]$$
$$z = t(x)$$

The former is the experimental, the latter the control group. t represents the effects of confounding variables, constant across treatment levels. If $y \neq z$ is significant there is a treatment effect. Given dissemination of research findings $E(y)$ (as expected values of y) during the evaluation cycle, the rational manager reacts by changing Φ to Φ^*, thus on Stufflebeam's argument, falsifying the evaluation report.

Introduce a reaction function R describing the rational manager's response:

$$\Phi^* = R\Phi \,.$$

By the contractive mapping theorem, in a metric space (Φ, d), for any reactions

$$R\Phi_1, \; R\Phi_2, \text{ if}$$
$$d(R\Phi_1, R\Phi_2) \leq kd(\Phi_1, \Phi_2)$$

where $o < k < 1$, then R has a unique fixed point Φ_o such that $\Phi_o^* = \Phi_o$ (2). This is the stable value of the treatment, where the manager will cease to modify the project in response to evaluative feedback. It is also the correct evaluative judgment, so that

$$y = E(y) = \Phi_o(x).$$

Hence we see that both the classical experimental design and treatment variance during the evaluation cycle, as a function of feedback, can be constituents of a general evaluation theory.

REFERENCES

1. Stufflebeam, Daniel L., "Use and Abuse of Evaluation in Title III" Theory into Practice, Vol. VI: 3 (1967), p. 128.

2. Kolmogorov, A. N. and Fomin, S. V., Elements of the Theory of Functions and Functional Analysis, Vol. I (New York: Graylock Press, 1957), p. 43.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 2, Winter 1968)

COEFFICIENT ALPHA, TEST RELIABILITY, AND HETEROGENEITY OF SCORE DISTRIBUTIONS

DONALD W. ZIMMERMAN
Carleton University

GRAHAM J. BURKHEIMER
East Carolina University

ABSTRACT

From concepts which refer only to observed scores and which allow the parameters of score distributions over repeated measurements on a given person to differ from person to person, necessary and sufficient conditions under which *coefficient alpha* equals test reliability are derived. The result clarifies the relation of this quantity to the Kuder-Richardson formula 20, to the KR 21, to the Spearman-Brown formula, and to Lord's item sampling model.

SINCE THE original derivation by Kuder and Richardson (6) of the formula which Cronbach designated as *coefficient alpha* the assumptions required have been more fully explicated (5, 4, 3, 2, 8). Derivations have been based on the classical test theory model in which true scores and error scores are uncorrelated, in which error scores on parallel tests are uncorrelated, and in which the parameters of error score distributions over repeated measurements on a person are the same for all persons. The purpose of this paper is to present a derivation of *coefficient alpha* somewhat different from those which have appeared in the literature, which reveals necessary and sufficient conditions under which this quantity is equal to the reliability of the test, and which indicates that the formula is valid under certain conditions where the classical test theory model is not applicable.

NOTATION AND DEFINITIONS

A notation similar to one introduced by Guttman (4) will be used in this paper. The subscript *g*, taking on values from 1 to H, will refer to repeated measurements on a given person. The subscript *j*, taking on values from 1 to K, will refer to persons. The subscript *i*, taking on values from 1 to N, will refer to the parts of a test. Expectations, variances, and covariances will be denoted by E, σ^2, and COV, with subscripts indicating the variables over which these values are taken.

We consider a set of pairs of independent repeated measurements, which are not yet identified with any empirical testing procedure. One member of each pair is denoted by X_{gji}, the other by X'_{gji}. Total test scores are X_{gj} and X'_{gj}.

The quantity X_{gj} is considered to be a random variable having expectation over repeated measurements, as the number of measurements increases without limit given by

$$\underset{g}{E}(X_{gj}) = \frac{\sum\limits_{g=1}^{H} X_{gj}}{H}$$

and variance given by

$$\underset{g}{\sigma^2}(X_{gj}) = \underset{g}{E}(X^2_{gj}) - \left[\underset{g}{E}(X_{gj})\right]^2 .$$

Similar expressions define $\underset{g}{E}(X'_{gj})$ and $\underset{g}{\sigma^2}(X'_{gj})$.

The quantities $\underset{g}{E}(X_{gji})$, $\underset{g}{E}(X'_{gji})$, $\underset{g}{\sigma^2}(X_{gji})$ and $\underset{g}{\sigma^2}(X'_{gji})$ refer to the expectations and variances of part-test scores for person *j* over repeated measurements.

Total observed variance is

$$\underset{g,j}{\sigma^2}(X_{gj}) = \underset{g,j}{E}(X^2_{gj}) - \left[\underset{g,j}{E}(X_{gj})\right]^2, \quad \text{where}$$

$$\underset{g,j}{E}(X_{gj}) = \frac{\sum\limits_{g=1}^{H}\sum\limits_{j=1}^{K} X_{gj}}{HK}.$$

A corresponding definition can be written for $\underset{g,j}{\sigma^2}(X'_{gj})$.

Similar expressions also define total observed part-test variances, $\underset{g,j}{\sigma^2}(X_{gji})$ and $\underset{g,j}{\sigma^2}(X'_{gji})$.

Total observed covariance of pairs of repeated measurements is given by

$$\underset{g,j}{\mathrm{COV}}(X_{gj}, X'_{gj}) = \underset{g,j}{E}(X_{gj}\ X'_{gj}) - \underset{g,j}{E}(X_{gj})\ \underset{g,j}{E}(X'_{gj}).$$

The reliability of a test is defined as follows:

$$\rho = \frac{\underset{g,j}{\mathrm{COV}}(X_{gj}, X'_{gj})}{\underset{g,j}{\sigma}(X_{gj})\ \underset{g,j}{\sigma}(X'_{gj})}.$$

The reliability of a part-test, ρ_i, is defined in the same way, in terms of X_{gji} and X'_{gji}.

It is assumed that the probability distributions of X_{gj} and X'_{gj} are the same for any person $\underline{j}$, such that

$$\underset{g}{E}(X_{gj}) = \underset{g}{E}(X'_{gj}) \text{ and } \underset{g}{\sigma^2}(X_{gj}) = \underset{g}{\sigma^2}(X'_{gj}).$$

It is not assumed that these probability distibutions are the same for different persons (12, in press). The same concept applies to the part-test scores, X_{gji} and X'_{gji}. It follows from the above that

$$\underset{g,j}{\sigma^2}(X_{gj}) = \underset{g,j}{\sigma^2}(X'_{gj}),\ \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gj})\right] = \underset{j}{\sigma^2}\left[\underset{g}{E}(X'_{gj})\right],$$

and $\rho\left[\underset{g}{E}(X_{gj}), \underset{g}{E}(X'_{gj})\right] = 1$, where $\rho\left[\underset{g}{E}(X_{gj}), \underset{g}{E}(X'_{gj})\right]$ is the product-moment correlation of $\underset{g}{E}(X_{gj})$ and $\underset{g}{E}(X'_{gj})$ for these K persons. Likewise, the same relationships hold for the part-test scores.

The quantity $\underset{g,j}{\mathrm{COV}}(X_{gj}, X'_{gj})$ is the sum of two quantities,

$$\underset{g,j}{\mathrm{COV}}(X_{gj}, X'_{gj}) = \underset{j}{E}\left[\underset{g}{\mathrm{COV}}(X_{gj}, X'_{gj})\right] + \underset{j}{\mathrm{COV}}\left[\underset{g}{E}(X_{gj}), \underset{g}{E}(X'_{gj})\right].$$

Since repeated measurements are independent, it follows that $\underset{g}{\mathrm{COV}}(X_{gj}, X'_{gj})$ is zero for all $\underline{K}$ persons. Therefore,

$$\underset{g,j}{\mathrm{COV}}(X_{gj}, X'_{gj}) = \underset{j}{\mathrm{COV}}\left[\underset{g}{E}(X_{gj}), \underset{g}{E}(X'_{gj})\right] = \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gj})\right] = \underset{j}{\sigma^2}\left[\underset{g}{E}(X'_{gj})\right].$$

The quantity $\underset{g,j}{\sigma^2}(X_{gj})$ is also the sum of two quantities,

$$\underset{g,j}{\sigma^2}(X_{gj}) = \underset{j}{E}\left[\underset{g}{\sigma^2}(X_{gj})\right] + \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gj})\right].$$

Using these results, together with the equalities established above for pairs of measurements, and making substitutions in the definition of reliability, we arrive at the result

$$(1) \qquad \rho = \frac{\underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gj})\right]}{\underset{g,j}{\sigma^2}(X_{gj})} = 1 - \frac{\underset{j}{E}\left[\underset{g}{\sigma^2}(X_{gj})\right]}{\underset{g,j}{\sigma^2}(X_{gj})}.$$

An equation for the reliability of a part-test score can be derived in the same way. We have

$$\rho_i = \frac{\underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]}{\underset{g,j}{\sigma^2}(X_{gji})} = 1 - \frac{\underset{j}{E}\left[\underset{g}{\sigma^2}(X_{gji})\right]}{\underset{g,j}{\sigma^2}(X_{gji})}.$$

Alternatively, if equation (1) is taken as a definition of reliability, as done by Guttman (4), the definition given in the present paper can be derived as a theorem.

The symbol $\underline{\alpha}$ will denote the quantity

$$\frac{N}{N-1}\left[1 - \frac{\sum\limits_{i=1}^{N} \underset{g,j}{\sigma^2}(X_{gji})}{\underset{g,j}{\sigma^2}(X_{gj})}\right].$$

We wish to establish necessary and sufficient conditions under which $\alpha = \rho$, using only the assumptions made in the present section. It should be noted that the definitions above refer only to observed scores, X_{gji}, that no assumptions concerning true and error components of scores and their correlations have been made, and that the parameters of score distributions over repeated measurements may differ from person to person.

DERIVATION OF COEFFICIENT ALPHA

Total observed variance can be expressed as the sum of variances and covariances of part-test scores as follows:

$$(2) \qquad \underset{g,j}{\sigma^2}(X_{gj}) = \sum_{i=1}^{N} \underset{g,j}{\sigma^2}(X_{gji}) + \sum_{\substack{i=1 \\ i \neq m}}^{N} \sum_{m=1}^{N} \underset{g,j}{\mathrm{COV}}(X_{gji}, X_{gjm}),$$

where the subscripts i and m refer to pairs of part-tests. Variance of $\underset{g}{E}(X_{gj})$ can likewise be written

$$(3)\quad \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gj})\right] = \sum_{i=1}^{N} \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right] + \sum_{\substack{i=1 \\ i \neq m}}^{N} \sum_{m=1}^{N} \underset{j}{\mathrm{COV}}\left[\underset{g}{E}(X_{gji}), \underset{g}{E}(X_{gjm})\right] .$$

Since $\underset{g,j}{\mathrm{COV}}(X_{gji}, X_{gjm}) = \underset{j}{E}\left[\underset{g}{\mathrm{COV}}(X_{gji}, X_{gjm})\right] + \underset{j}{\mathrm{COV}}\left[\underset{g}{E}(X_{gji}), \underset{g}{E}(X_{gjm})\right]$

and since $\underset{g}{\mathrm{COV}}(X_{gji}, X_{gjm})$ is zero for all j,

$$\underset{g,j}{\mathrm{COV}}(X_{gji}, X_{gjm}) = \underset{j}{\mathrm{COV}}\left[\underset{g}{E}(X_{gji}), \underset{g}{E}(X_{gjm})\right] .$$

Subtracting equation (3) from equation (2) and using the relationships

$$\underset{g,j}{\sigma^2}(X_{gj}) = \underset{j}{E}\left[\underset{g}{\sigma^2}(X_{gj})\right] + \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gj})\right] \text{ and}$$

$$\underset{g,j}{\sigma^2}(X_{gji}) = \underset{j}{E}\left[\underset{g}{\sigma^2}(X_{gji})\right] + \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right], \text{ we obtain}$$

$$\underset{j}{E}\left[\underset{g}{\sigma^2}(X_{gj})\right] = \sum_{i=1}^{N} \underset{g,j}{\sigma^2}(X_{gji}) - \sum_{i=1}^{N} \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]$$

$$= \sum_{i=1}^{N} \underset{j}{E}\left[\underset{g}{\sigma^2}(X_{gji})\right] .$$

Then, from equation (1)

$$(4)\quad \rho = 1 - \frac{\sum_{i=1}^{N} \underset{j}{E}\left[\underset{g}{\sigma^2}(X_{gji})\right]}{\underset{g,j}{\sigma^2}(X_{gj})}$$

$$= 1 - \frac{\sum_{i=1}^{N} \underset{g,j}{\sigma^2}(X_{gji})}{\underset{g,j}{\sigma^2}(X_{gj})} + \frac{\sum_{i=1}^{N} \underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]}{\underset{g,j}{\sigma^2}(X_{gj})} .$$

Since $\underset{i,j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right] = \underset{i}{E}\left\{\underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} + \underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\}$

and $\underset{i,j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right] = \underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} + \underset{j}{\sigma^2}\left\{\underset{i}{E}\left[\underset{g}{E}(X_{gji})\right]\right\}$,

it follows that

$$\underset{i}{E}\left\{\underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} = \underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} + \underset{j}{\sigma^2}\left\{\underset{i}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} - \underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} .$$

Substituting this result in (4) gives

$$(5)\quad \rho = 1 - \frac{\sum_{i=1}^{N} \underset{g,j}{\sigma^2}(X_{gji})}{\underset{g,j}{\sigma^2}(X_{gj})} + \frac{N\underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\}}{\underset{g,j}{\sigma^2}(X_{gj})} + \frac{N\underset{j}{\sigma^2}\left\{\underset{i}{E}\left[\underset{g}{E}(X_{gji})\right]\right\}}{\underset{g,j}{\sigma^2}(X_{gj})} - \frac{N\underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\}}{\underset{g,j}{\sigma^2}(X_{gj})}$$

Since

$$\underset{i}{E}\left[\underset{g}{E}(X_{gji})\right] = \frac{\underset{g}{E}(X_{gj})}{N}, \quad \underset{j}{\sigma^2}\left\{\underset{i}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} = \frac{\underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gj})\right]}{N^2} .$$

Therefore,

$$\frac{N\underset{j}{\sigma^2}\left\{\underset{i}{E}\left[\underset{g}{E}(X_{gji})\right]\right\}}{\underset{g,j}{\sigma^2}(X_{gj})} = \frac{\underset{j}{\sigma^2}\left[\underset{g}{E}(X_{gj})\right]}{N\underset{g,j}{\sigma^2}(X_{gj})} = \frac{\rho}{N} .$$

Substituting $\frac{\rho}{N}$ in (5) and solving for ρ leads to the following result

$$(6)\quad \rho = \frac{N}{N-1}\left[1 - \frac{\sum_{i=1}^{N} \underset{g,j}{\sigma^2}(X_{gji})}{\underset{g,j}{\sigma^2}(X_{gj})} + \frac{N\underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} - N\underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\}}{\underset{g,j}{\sigma^2}(X_{gj})}\right] .$$

which except for the right-hand term, is the same as α .

We can write

$$\alpha = \rho \Leftrightarrow \underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} = \underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} .$$

In other words, coefficient alpha is equal to the reliability of the test if and only if, in a population of repeated measurements, the mean over persons of the variances over part-tests of the mean part-test scores over repeated measurements is equal to the variance over part-tests of the means over persons of the mean part-test scores over repeated measurements.

We consider now conditions under which these quantities are equal. If, for all i, j, $\underset{g}{E}(X_{gji}) = A_i + B_j$, where A_i varies over i but is the same for all j and B_j varies over j but is the same for all i, then

$$\underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} = \underset{j}{E}\left\{\underset{i}{\sigma^2}\left[A_i + B_j\right]\right\}$$

$$= \underset{j}{E}\left[\underset{i}{\sigma^2}(A_i)\right]$$

$$= \underset{i}{\sigma^2}(A_i) .$$

Also,

$$\underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} = \underset{i}{\sigma^2}\left\{\underset{j}{E}\left[A_i + B_j\right]\right\}$$

$$= \underset{i}{\sigma^2}\left[A_i + \underset{j}{E}(B_j)\right]$$

$$= \underset{i}{\sigma^2}(A_i) \quad .$$

That is, each $\underset{g}{E}(X_{gji})$ value over the entire N x K matrix is the sum of a row effect, which is invariant over columns, and column effect, which is invariant over rows. A verbal expression of this condition is that the score of each person on each part-test is determined by two "effects" -- "ability" and "test difficulty" (or "item difficulty"), which do not interact.

Also, it follows from the above that for all i and m

$$\underset{g,j}{\mathrm{COV}}(X_{gji}, X_{gjm}) = \underset{g,j}{\mathrm{COV}}(X_{gji}, X'_{gji}) = \underset{g,j}{\mathrm{COV}}(X_{gjm}, X'_{gjm}),$$

$$\underset{g,j}{\sigma^2}(X_{gji}) = \underset{g,j}{\sigma^2}(X_{gjm}),$$

$$\rho_i = \rho_m = \rho(X_{gji}, X_{gjm}) \text{ , and}$$

$$\rho\left[\underset{g}{E}(X_{gji}) \text{ , } \underset{g}{E}(X_{gjm})\right] = 1 \quad .$$

In other words, the variances and covariances of all part-tests are equal, the intercorrelations among all parts are equal and equal to the reliability of each part, and the correlation between the $\underset{g}{E}(X_{gji})$ values is unity for all pairs of part tests.

In the classical test theory model

$$X_{gj} = T_j + E_{gj} \text{ and}$$

$$T_j = \lim_{H \to \infty} \frac{\sum_{g=1}^{H} X_{gj}}{H} \text{ , where } T_j \text{ and } E_{gj} \text{ are true}$$

and error components of scores. In the present derivation the quantity $\underset{g}{E}(X_{gj})$ has not been identified with the concept of true score. If, however, $\underset{g}{E}(X_{gj})$ is a linear function of T_j, the conditions stated by Novick and Lewis (8) are applicable and hold even when true and error components of scores are correlated and when the distributions of scores over repeated measurements are different from person to person (1, 9, 10, 11). That is, even in this case a necessary and sufficient condition for coefficient alpha to equal test reliability is that the part-tests be essentially τ - equivalent, as defined by Novick and Lewis.

From a single administration of a test an estimate of α is provided by the quantity

$$\frac{N}{N-1}\left[1 - \frac{\sum_{i=1}^{N} s^2(X_i)}{s^2(X)}\right],$$

where $s^2(X_i)$ is the variance of scores on part i over all K persons on a single administration of a test and $s^2(X)$ is the variance of total scores over all K persons on a single adminstration of a test.

If conditions for α are met, the covariance between pairs of items within the test equals the covariance between pairs of items on parallel tests. Further, if conditions for the KR 21 are met, for any given person the variance of scores over items equals the variance of scores on any item over repeated testings--that is

$$\underset{g,i}{\sigma^2} X_{gji} = \underset{g}{\sigma^2} X_{gji} \text{ , for all i, j.}$$

When individual items are considered α is equivalent to the Kuder-Richardson formula 20. That formula, then, is equal to the reliability of a test if and only if test items satisfy the above conditions.

EQUIVALENCE OF COEFFICIENT ALPHA AND THE SPEARMAN-BROWN FORMULA

The conditions under which the reliability of a composite test of N parts is equal to coefficient alpha are identical to the conditions under which the reliability of a test lengthened N times is given by the Spearman-Brown formula. The latter condition is sometimes expressed by saying that the parts added to the original test must all be parallel. But under the usual definition of parallelism this is a sufficient condition for coefficient alpha, while the extension of the concept of parallelism made by Novick and Lewis is a necessary and sufficient condition. In the case in which N refers to items the conditions under which the reliability of a test of N items is given by the Kuder-Richardson formula 20 are identical to the conditions under which the reliability of a 1-item test lengthened N times is given by the Spearman-Brown formula.

The relationship can be seen by letting ρ_i be the reliability of one part-test, or one item, before the test is lengthened. If the above conditions are met, $\sum_{i=1}^{N} \underset{g,j}{\sigma^2}(X_{gji}) = N \underset{g,j}{\sigma^2}(X_{gji})$, and we can substitute in α an equivalent expression in terms of part-test variances and covariances and write

$$(9) \qquad \rho = \frac{N}{N-1}\left[1 - \frac{\underset{g,j}{\sigma^2}(X_{gji})}{N \underset{g,j}{\sigma^2}(X_{gji})\left[1 + (N-1)\rho_i\right]}\right] = \frac{N\rho_i}{1 + (N-1)\rho_i} \quad .$$

These conditions can be summarized as follows. A sufficient condition for coefficient alpha to equal the reliability of a composite test of N parts is that the parts be parallel, in the sense in which this term has been defined in test theory. A necessary and sufficient condition for coefficient alpha to equal the reliability of a composite test is that the parts be essentially τ- equivalent, as defined by Novick and Lewis, provided the $\underset{g}{E}(X_{gji})$ values are linearly related to true scores. A sufficient condition for the Spearman-Brown formula to equal the reliability of a test lengthened N times is that the added parts be parallel. Finally, a necessary and sufficient condition for the Spearman-Brown formula to equal the reliability of a test lengthened N times is that the added parts be essentially τ-equivalent. Still another way of expressing the condition for the Spearman-Brown formula is that in lengthening the test the expected observed score of each person on each of the added parts, $\underset{g}{E}(X_{gji})$, is the sum of two components, A_i and B_j, where A_i, referring to the difficulty of the part, is the same for all persons and B_j, referring to the ability of the person, is the same for all parts.

RELATION OF THE KUDER-RICHARDSON FORMULA 20 TO THE KUDER-RICHARDSON FORMULA 21

Assume that the conditions for $\alpha = \rho$ are met. Then, for a test consisting of dichotomous items

$$\alpha = \rho = \frac{N}{N-1}\left[1 - \frac{\sum_{i=1}^{N} \underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\left\{1-\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\}}{\underset{g,j}{\sigma^2}(X_{gj})}\right].$$

$$= \frac{N}{N-1}\left[1- \frac{\underset{g,j}{E}(X_{gj})\left[N-\underset{g,j}{E}(X_{gj})\right]}{N\,\underset{g,j}{\sigma^2}(X_{gj})} + \frac{N\underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\}}{\underset{g,j}{\sigma^2}(X_{gj})}\right].$$

When the right-hand term is zero, this result reduces to the Kuder-Richardson formula 21. In other words, the condition required for the KR 20 formula to equal the reliability of the test is that

$$\underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} = \underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\},$$

and the condition required for the KR 21 is that

$$\underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} = \underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} = 0 .$$

Or, for a test consisting of dichotomous items, the condition for the KR 20 formula is that for all i, j $\underset{g}{E}(X_{gji}) = A_i + B_j$, and the condition for the KR 21 formula is that for all i, j $\underset{g}{E}(X_{gji}) = B_j$. Equation (6) gives the reliability of a test without restriction on the $\underset{g}{E}(X_{gji})$ values or on the nature of the parts.

If the KR 21 formula is applicable, an estimate of reliability from a single administration of a test is provided by

$$\frac{N}{N-1}\left[1 - \frac{\bar{X}(N-\bar{X})}{Ns^2(X)}\right],$$

where $\bar{X}$ is the mean of total scores over K persons on a single administration of a test.

DERIVATION OF COEFFICIENT ALPHA FROM ITEM SAMPLING MODEL

An interpretation of the Kuder-Richardson formula 21 can be based on the model presented by Lord for the sampling of test items from a population of items. If Lord's expression for the standard error of measurement of a person's score,

$$\frac{X_{gj}(N - X_{gj})}{N - 1},$$

is identified with error variance and if its arithmetic mean is taken over K persons and substituted in the classical equation, $\rho = 1 - \frac{\sigma_e^2}{\sigma_X^2}$,

the KR 21 formula is obtained (7). The KR 21 formula, therefore, can be interpreted as an equation for test reliability for the case in which the observed score distribution over repeated measurements is identified with variations in score resulting from item sampling.

It should be noted that the requirement that

$$\underset{j}{E}\left\{\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]\right\} = \underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} = 0$$

is implicit in this model. From the concept of item sampling employed by Lord it follows that, as H increases without limit, for any j, $\underset{i}{E}\left[\underset{g}{E}(X_{gji})\right] = \underset{g}{E}(X_{gji})$ for all i, such that $\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right]$ becomes zero for all j. Further, $\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]$ becomes the same for all i, such that $\underset{i}{\sigma^2}\left\{\underset{j}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} = 0$. Also note that the distribution over independent repeated measurements of scores on dichotomous items is necessarily binomial, apart from any identification of the set of repeated measurements with an empirical sampling process.

Let us assume now that the N items of a test are drawn from N distinct populations and that the values of $\underset{g}{E}(X_{gji})$ for person j may differ. Then, the distribution of observed scores over repeated measurements is given by a Poisson sequence of trials,

$$\underset{g}{\sigma^2}(X_{gj}) = N \underset{i}{E}\left[\underset{g}{E}(X_{gji})\right]\left\{1-\underset{i}{E}\left[\underset{g}{E}(X_{gji})\right]\right\} - N\underset{i}{\sigma^2}\left[\underset{g}{E}(X_{gji})\right].$$

Finding the arithmetic mean of this quantity over all K persons and substituting the result in equation (1), we obtain (6). Therefore, equation (6) gives test reliability for the case in which the N items of a test for K persons are sampled from NK distinct populations of items. Given the same restrictions mentioned in the last section the KR 20 formula and the KR 21 formula, respectively, follow from the item-sampling interpretation. However, the value of $\sigma^2_g(X_{gj})$ for test scores consisting of the sum of scores on dichotomous items is given by the above expression and the value of ρ is given by equation (6), apart from any empirical identification of the set of repeated measurements.

SUMMARY

Beginning by considering distributions of observed scores over independent repeated measurements, the following was proved:

1. Coefficient alpha is equal to the reliability of a test if and only if, in a population of repeated measurements, the mean over persons of the variances over part-tests of the mean part-test scores over repeated measurements is equal to the variance over part-tests of the means over persons of the mean part-test scores over repeated measurements. This conclusion holds even if observed scores are regarded as the sum of components which are linearly correlated and even if the distributions of scores over repeated measurements are different from person to person. This condition is equivalent to the assumption that the $\underset{g}{E}(X_{gji})$ value of each person on each part-test is the sum of two effects--one, "ability," which is invariant over items, and the other, "test difficulty," which is invariant over persons--i. e., that $\underset{g}{E}(X_{gji}) = A_i + B_j$, for all i, j.

2. The Kuder-Richardson formula 20 is equal to the reliability of a test if and only if test items satisfy the above equality. The Kuder-Richardson formula 21 is equal to the reliability of a test if and only if both of the above quantities are zero.

3. The conditions under which the reliability of a composite test of N parts is equal to coefficient alpha are identical to the conditions under which the reliability of a test lengthened N times is given by the Spearman-Brown formula.

4. Both the KR 20 and the KR 21 can be interpreted as equations for test reliability for the case in which for each person the N items of a test are sampled from populations defined by the conditions given above. Equation (6) gives test reliability for the case in which the N items of a test for K persons are sampled from NK populations of items without restrictions.

FOOTNOTE

1. This research was supported by the U. S. Office of Education under grant OEC 2-7-068209-0389.

REFERENCES

1. Burkheimer, G. J., Zimmerman, D. W., and Williams, R. H. "The Maximum Reliability of a Multiple-choice Test as a Function of Number of Items, Number of Choices, and Group Heterogeneity," Journal of Experimental Education," XXXV, (1967). pp. 89-94.
2. Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, XVI, (1951), pp. 297-334.
3. Gulliksen, H. Theory of Mental Tests (New York: Wiley, 1950).
4. Guttman, L. " A Basis for Analyzing Test-retest Reliability," Psychometrika, X (1945), pp. 255-282.
5. Jackson, R. W. B. and Ferguson, G. A., "Studies on the Reliability of Tests," Bulletin No. 12, Department of Educational Research, University of Toronto, (1941).
6. Kuder, G. F. and Richardson, M. W., "The Theory of the Estimation of Test Reliability," Psychometrika, II, (1937) pp. 151-160.
7. Lord, F. M. "Sampling Fluctuations Resulting from the Sampling of Test Items," Psychometrika, XX, (1955), pp. 1-22.
8. Novick, M. R. and Lewis, C. "Coefficient Alpha and the Reliability of Composite Measurements," Psychometrika, XXXII, (1967), pp. 1-13.
9. Williams, R. H. and Zimmerman, D. W. "An Extension of the Rulon Formula for Test Reliability: The Case of Correlated True and Error Components of Scores," Journal of Experimental Education, XXXVI, No. 3 (Spring, 1968), pp. 94-96.
10. Zimmerman, D. W. and Williams, R. H. "Chance Success Due to Guessing and Non-independence of True Scores and Error Scores in Multiple-choice Tests: Computer Trials With Prepared Distributions," Psychological Reports, XVII, (1965) pp. 159-165.
11. Zimmerman, D. W. and Williams, R. H. "Generalization of the Spearman-Brown Formula for Test Reliability: The Case of Non-independence of True Scores and Error Scores," British Journal of Mathematical and Statistical Psychology, XIX, (1966), pp. 271-274.
12. Zimmerman, D. W., Williams, R. H. and Burkheimer, G. J. "Dependence of Test Reliability upon Heterogeneity of Individual and Group Score Distributions," Educational and Psychological Measurement, XXVIII, (1968), pp. 41-46.

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION

(Act of October 23, 1962: Section 4369, Title 39, United States Code)

Publisher: File two copies of this form with your postmaster.
Postmaster: Complete verification on page 2

Form Approved.
Budget Bureau No. 46-R029

1 DATE OF FILING
10-24-68

2 TITLE OF PUBLICATION
JOURNAL OF EXPERIMENTAL EDUCATION

3 FREQUENCY OF ISSUE
QUARTERLY

4 LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, city, county, state, ZIP code)*
2101 Sherman Avenue - Madison Dane Wisconsin 53701

5 LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*
2101 Sherman Avenue - Madison Dane Wisconsin 53701

6 NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER *(Name and address)*
Wilson Thiede - 4825 Bayfield Terrace - Madison, Wisconsin

EDITOR *(Name and address)*
John Schmid - Colorado State College - Greeley, Colorado

MANAGING EDITOR *(Name and address)*
Norma C. Camp- Dowd Road - Waunakee, Wisconsin

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)* DEMBAR EDUCATIONAL RESEARCH SERVICES, INC.

NAME	ADDRESS
	2101 Sherman Avenue Madison, Wisconsin
Walter A. Frautschi	29 Fuller Drive - Madison, Wisconsin
Dorothy J. Frautschi	29 Fuller Drive - Madison, Wisconsin
John J. Frautschi	221 Lakewood Blvd. - Madison, Winconsin
Walter J. Frautschi	3206 Lake Mendota Dr.- Madison, Wisconsin

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

NAME	ADDRESS
None	

9. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES *(Section 132.122, Postal Manual)*
(Check one)

The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes ☐ Have not changed during preceding 12 months ☐ Have changed during preceding 12 months *(If changed, publisher must submit explanation of change with this statement.)*

10. EXTENT AND NATURE OF CIRCULATION	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NUMBER OF COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
A. TOTAL NO. COPIES PRINTED *(Net Press Run)*	2,225	2,300
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES		
2. MAIL SUBSCRIPTIONS	1,901	2,059
C. TOTAL PAID CIRCULATION	1,901	2,059
D. FREE DISTRIBUTION *(including samples)* BY MAIL, CARRIER OR OTHER MEANS	57	64
E. TOTAL DISTRIBUTION *(Sum of C and D)*	1,958	2,123
F. OFFICE USE, LEFT-OVER, UNACCOUNTED, SPOILED AFTER PRINTING	267	177
G. TOTAL *(Sum of E & F—should equal net press run shown in A)*	2,225	2,300

I certify that the statements made by me above are correct and complete. ▶ *(Signature of editor, publisher, business manager, or owner)*
Arnold A. Caucutt, Business Mgr.

POD Form 3526 May 1968

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

ATTITUDES OF ELEMENTARY SCHOOL TEACHERS TOWARD COUNSELING AND GUIDANCE IN THE ELEMENTARY SCHOOLS[1]

FREDERICK J. AXELBERD

University of Georgia

ABSTRACT

This study was designed to determine if teachers employed in elementary schools that have maintained a counseling and guidance program would express more favorable attitudes toward counseling and guidance in the elementary schools than teachers employed in elementary schools which have not provided this service. Attitude scales were distributed to 566 teachers and 388 were returned completed. Analysis of Variance revealed a trend in which teachers employed in elementary schools that have maintained a counseling and guidance program for at least 2 years expressed more favorable attitudes than teachers employed in elementary schools not providing these services. In addition, significant differences between counties were found, which indicated that other factors influenced the teacher attitudes.

THE PRESENT study is concerned with the expressed attitudes of elementary school teachers toward counseling and guidance in the elementary schools. The motivation for conducting such a project is based on the contention, supported by research and theory (4, 6, 1, 5), that the attitudes of teachers are most relevant, significant, and valuable to counselors and educational administrators as they work together, to organize, to develop, to maintain, and to improve school programs.

The study was designed to answer the following question:

> Will teachers employed in elementary schools that have maintained a counseling and guidance program for at least 2 years express more favorable attitudes toward counseling and guidance than teachers employed in elementary schools which have not provided this service?

METHODOLOGY

The teachers in the selected schools were asked to complete two attitude scales. The instruments administered were modified versions of the Counseling Attitude Scale and the Scale of Attitudes Toward School Guidance (for a comprehensive analysis of these instruments see 4, 3). Tests of reliability yielded coefficients of correlation ranging upward from .86 indicating high reliability. The scores on the two scales yielded a .65 coefficient of correlation.

The 388 teachers who responded represented 68.55 percent of the 566 teachers employed in the participating schools. Seven elementary schools with counseling and guidance programs for at least 2 years were selected for study and designated project schools. Two control schools were selected, one from the same community and one from the same county, for each project school. An attempt was made to match the project schools with control schools as to size, number of teachers, socio-economic level of the community, and, in the case of the community control schools, geographic proximity. Some teachers in the control schools had contact with counselors who were providing guidance services to several schools on a part-time schedule; however, they did not have the experience of working with a counselor involved only with their school.

A general description of the three counties (all in the same state) in which the project schools are located is presented:

> County One is located on the east coast and is dominated by commercial and industrial interests, especially shipping. The teachers of this county have recently invoked "professional sanctions" in response to the overcrowded condition of many schools and low salaries.
>
> County Two is located on the east coast and is dominated by resort and tourist interests. It is a major popu-

lation center of the state. School conditions in the county are generally good, although existing physical facilities are reaching their capacity for new students. Teacher salaries are the highest of the counties sampled. Six schools participated from this county.

County Three is located on the east coast and is not dominated by any one group. It has a prospering tourist industry, a productive agriculture, and light industry. This is a moderately populated county. Although three project schools are situated in this county, a community control school was available for only one of them. Teachers in the other two schools participated in the survey and their responses are included in analyses unrelated to the comparison of project schools with their community control schools.

In addition, elementary schools, located in rural counties were sampled until the number of teachers selected in these schools were comparable to the number chosen from the seven project schools. Eight rural schools from four different counties were involved. None of these rural schools had employed a full-time counselor.

An analysis of variance, corrected for cells with unequal frequencies, was made to determine whether significant variances existed between the project and control schools' respondents. A similar analysis was undertaken to determine whether the variance between counties 1, 2, 3, and rural was statistically significant.

RESULTS AND DISCUSSION

Teachers from project schools expressed significantly more favorable attitudes than teachers from community control schools; however, this finding needs to be qualified because the attitude scores, when grouped according to county, exhibited significant differences between the counties sampled. This indicated that factors or conditions within a school

TABLE 1

RESULTS OF ANALYSIS OF VARIANCE

Source of Variance	S. S.	df.	F
Project and Control Schools			
Counseling Attitude Scale			
Between Groups	1, 207	3	2. 94*
Within Groups	47, 578	348	
Total	48, 785	351	
Scale of Attitudes Toward School Guidance			
Between Groups	666	3	1. 06
Within Groups	77, 655	348	
Total	78, 321	351	
Counties 1, 2, 3, and Rural			
Counseling Attitude Scale			
Between Groups	1, 664	3	3. 49*
Within Groups	55, 479	348	
Total	57, 143	351	
Scale of Attitudes Toward School Guidance			
Source of Variance			
Between Groups	1, 511	3	2. 42
Within Groups	72, 716	348	
Total	74, 227	351	

* p <.05

TABLE 2

DIFFERENCES AND t VALUES FOR MEAN SCORES

Project vs. Controls	t_1	t_2
Project vs. Community Controls	2.84*	1.46
Project vs. County Controls	.86	.23
Project vs. Rural Schools	1.77	.99
Community Controls vs. County Controls	2.31*	1.32
Community Controls vs. Rural Schools	.65	.49
County Controls vs. Rural Schools	1.15	.81
County		
County 1 vs. County 2	3.18**	2.10*
County 1 vs. County 3	2.36*	2.73**
County 1 vs. Rural Schools	1.14-	.64
County 2 vs. County 3	.73	1.04
County 2 vs. Rural Schools	1.72	1.38
County 3 vs. Rural Schools	1.05	2.10*

* $p < .05$
** $p < .01$

t_1 - Derived from Counseling Attitude Scale.

t_2 - Derived from Scale of Attitudes Toward School Guidance.

system, other than the existing counseling and guidance programs, exerted influence on the attitudes expressed by the teachers (see tables 1 and 2).

The difference that existed between the expressed attitudes of teachers from project and control schools, as well as the differences between counties 1, 2, 3, and rural, were different in the degree of favorability toward counseling and guidance in elementary schools. Since teachers from project schools expressed significantly more favorable attitudes (on the Counseling Attitude Scale) than teachers from community control schools, teaching in a school with a full-time elementary school counselor is related to more favorable teacher attitudes than teaching in an elementary school which does not have a counselor on its faculty.

Another factor was found to be related to teachers' expressed attitudes; county control respondents expressed significantly more favorable attitudes than community control respondents (on the Counseling Attitude Scale). The difference is accounted for by comparing the disproportional contribution of counties 1 and 2 to the number of teachers in these control groups. Highly significant differences, on both scales, in favor of county 2 over county 1 were found; since county 1 contributed twenty-eight teachers to the county controls and county 2, forty-four teachers, the influence of county 2 is greater than county 1 in the county control group. The more favorable mean scores of county control respondents reflects the positive influence of teachers from county 2 in this control group.

The differences between the mean scores of project and control schools (although the only significant difference was between the means of project and community controls), indicates an attitude trend in which teachers working in elementary schools maintaining a full-time counselor have more favorable attitudes toward counseling and guidance than teachers employed in schools having more limited or no counseling and guidance programs.

Similarly, the analysis of variance between counties revealed a significant trend. When the mean scores of county 2 on the Counseling Attitude Scale and Scale of Attitudes Toward School Guidance were compared to the mean scores of counties 1 and 3, the significant differences were in favor of county 2.

CONCLUSIONS AND IMPLICATIONS

1. Teachers employed in elementary schools that have maintained a counseling and guidance program for at least 2 years expressed more favorable attitudes toward counseling and guidance than teachers employed in elementary schools which have not provided these services. This

conclusion is qualified, in light of the significant differences obtained between the counties sampled. It is suggested that variance in teacher attitudes toward counseling and guidance in elementary schools may be related to factors other than the existing counseling and guidance program in a school.

2. The differences between the mean scores of the counties sampled were statistically significant and the implication for educators responsible for evaluating educational programs in more than one county is that the evaluator be cognizant of the diversity of variables within a school system that could effect teacher attitudes.

3. It is important to point out that approximately 30 percent of the teachers asked did not choose to participate in the study. This 30 percent may include teachers who exhibit a common variable or variables related to negative as well as positive attitudes toward counseling and guidance in elementary schools.

FOOTNOTE

1. This study is based on the author's doctoral dissertation (1967) supervised by Dr. Robert O. Stripling, Chairman and Professor, Department of Personnel Services, University of Florida.

REFERENCES

1. Aubrey, R. D., "The Effect of Counselors on the Reward System of Teachers," Personnel and Guidance Journal, 45:1017-1020, 1967.

2. Axelberd, F. J., "Attitudes of Elementary School Teachers Toward Counseling and Guidance in the Elementary Schools," unpublished EdD dissertation, University of Florida, 1967.

3. Barker, D. D., "Development of a Scale of Attitudes," Personnel and Guidance Journal, 44: 1077-1083, 1966.

4. Form, A. L., "The Construction of a Scale on Attitudes Toward Counseling," Journal of Counseling Psychology, 34: 96-102, 1955.

5. Greising, R. A., "A Pilot Program in Elementary School Counseling," Elementary School Guidance and Counseling, 1: 126-139, 1967.

6. Stewart, J. A., "Factors Influencing Teacher Attitudes Toward and Participation in Guidance Services," Personnel and Guidance Journal, 39: 729-734, 1961.

Important Books From DERS

ABILITY GROUPING IN THE PUBLIC SCHOOLS

By Walter R. Borg. Foreword by Philip Lambert. 1966

104 pages, $3

This monumental field study analyzes differences in the effects upon elementary, junior high, and high school pupils of an ability grouping system, reporting on a wide range of variables, and drawing significant conclusions.

"At last here is a volume on ability grouping that considers the results as well as the possibilities."—SCHOOL MANAGEMENT.

The Measurement and Prediction of
TEACHER EFFECTIVENESS

A. S. Barr, with Worcester, Abell, Beecher, Jensen, Peronto, Ringness, and Schmid. 1961

144 pages, $5

What are the criteria of teacher effectiveness? What are the pertinent theories of the nature and structure of human abilities? Are there discrepancies between teacher potential and performance? What are the personal and professional prerequisites to teacher effectiveness? Can good teachers be distinguished from poor teachers?
This milestone monograph presents an overview of 30 years of important research into teacher effectiveness, and offers such general observations and new hypotheses as appear to be supported by available data.

"Our all-time best seller."—WILSON THIEDE, Publisher.

ders

DEMBAR
EDUCATIONAL
RESEARCH
SERVICES, INC.

POST OFFICE BOX 1148 • MADISON, WISCONSIN 53701

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

A SYMBOLIC SYSTEM FOR CLASSIFYING MULTIVARIATE PROCEDURES

ANDREW R. BAGGALEY

University of Pennsylvania

ABSTRACT

Procedures are classified according to whether the dependent or independent variable sets consist of single or multiple variables, whether they are categorical or metric in nature, and whether or not they are statistically held constant.

MULTIVARIATE procedures are increasingly being used by investigators in education and the behavioral sciences. Nowadays it is a rare study that deals with the relation between only two variables. Multiscore tests and multiple ratings are widely employed. Therefore it seems useful to have available a concise symbolic classification to aid the investigator in selecting a procedure appropriate for his data.

Table 1 shows such a taxonomic scheme. Of course, Guttman's lambda (5:71-8), simple correlation, and the correlation ratio are not multivariate statistics, but they furnish a baseline against which the other, more elaborate techniques can be compared. The elaborations consist of replacing a single variable by a set of variables and partialing out variables or sets of variables in various combinations.

The third, fourth, and fifth procedures constitute an interesting group. Most textbooks deal with only the fifth, i. e., partial correlation. Let X_0 represent a criterion variable (e. g., collegiate grade point average), X_1 a predictor variable (e. g., score on a study habits inventory), and X_2 a concomitant variable (e. g., score on a scholastic aptitude test). The partial correlation is given by

$$r_{(0.2)\,(1.2)} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1 - r_{02}^2}\ \sqrt{1 - r_{12}^2}}$$

In this procedure the concomitant variable is partialed out of both the predictor and criterion, DuBois (4:208-10) has described an alternative technique that he calls "part correlation," in which the concomitant variable is partialed out of only the predictor. The formula is

$$r_{(0)\,(1.2)} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1 - r_{12}^2}}$$

Another possibility is to partial the concomitant variable out of only the criterion. The formula for this variant is

$$r_{(0.2)\,(1)} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1 - r_{02}^2}}$$

This statistic is here called "semipartial correlation," although the term is often used synonomously with part correlation. Notice that the three formulas differ only in their denominators. Semipartial correlation seems most appropriate for the example given above. It shows how much the students' study habits are related to the component of their grade point average that is independent of their scholastic aptitude. In Table 1 compare the paradigm for semipartial correlation with that for one-way analysis of covariance. They differ only in the nature of the predictor, which is categorical in the case of the latter. In our example we could replace score on the study habits inventory with the dichotomy of public school vs. private school graduate. Then we would be comparing the two types of school with regard to grade point average, with scholastic aptitude partialed out. Jennings (7) has presented a set of matrix formulas from which part and partial correlation can be derived when one or more variables are to be partialed out. A simple interchange of symbols in Jennings' formulas covers semipartial correlation.

Another interesting comparison is that between multivariate analysis of variance and discriminant analysis (3:60-3, 116-21). In both cases a categorical variable and a set of metric variables are involved. In multivariate analysis of variance the categorical variable usually represents a stimulus variable, i. e., one that is physically manipulated (1:2). The interest is in the presumed effect of this manipulation on a set of metric variables. Discriminant analysis is usually applied when both the categorical variable and the set of metric variables are observed as they occur in nature. For example, Hall (6) compared the mean profiles of scores on the Differential Aptitude Tests of groups of high school students who entered various occupational fields.

The reader may wonder at the absence of factor analysis from Table 1. In each of the paradigms in the table the variables are partitioned into two sets; one of the sets is regarded as being dependent on the other in some sense. In factor analysis all of the variables are treated as having the same status in a single set. One could follow Tatsuoka and Tiedeman (9), by allowing the set of dependent variables to be empty and thus include factor analysis. Another interesting generalization is to transform the categorical variable(s) into sets of dichotomous variables and thus regard fixed effects analysis of variance as a regression problem (8).

Study of the paradigms in Table 1 will show other interesting comparisons between the various multivariate procedures. Investigators will be glad to learn that half of these procedures can be handled by a single computer program, the MANOVA program of the University of Miami's Biometric Laboratory (2).

REFERENCES

1. Baggaley, Andrew R., Intermediate Correlational Methods, John Wiley and Sons, Inc., New York, 1964, 211 pp.
2. Clyde, Dean J., Cramer, Elliot M. and Sherin, Richard J., Multivariate Statistical Programs, University of Miami Biometric Laboratory, Coral Gables, 1966, 61 pp.
3. Cooley, William W. and Lohnes, Paul R., Multivariate Procedures for the Behavioral Sciences, John Wiley and Sons, Inc., New York, 1962, 211 pp.
4. DuBois, Philip H., An Introduction to Psychological Statistics, Harper and Row, New York, 1965, 530 pp.
5. Freeman, Linton C., Elementary Applied Statistics, John Wiley and Sons, Inc., New York, 1965, 298 pp.
6. Hall, Robert C., "Occupational Group Contrasts in Terms of the Differential Aptitude Tests: An Application of Multiple Discriminant Analysis," Educational and Psychological Measurement, 17:556-67, Winter 1957.
7. Jennings, Earl, "Matrix Formulas for Part and Partial Correlation," Psychometrika, 30:353-6, September 1965.
8. Jennings, Earl, "Fixed Effects Analysis of Variance by Regression Analysis," Multivariate Behavioral Research, 2:95-108, January 1967.
9. Tatsuoka, Maurice M. and Tiedeman, David V., "Statistics as an Aspect of Scientific Method in Research on Teaching," in N. L. Gage (Ed.), Handbook of Research on Teaching, Rand McNally and Co., Chicago, 1963.

TABLE 1

PARADIGMS FOR MULTIVARIATE PROCEDURES

C: a categorical variable (a single way of classifying people or objects)

C': a set of categorical variables

M: a metric (continuous) variable

M': a set of metric variables

$C = f(C)$: Guttman's lambda

$M_1 = f(M_2)$: simple (two-variable) correlation

$M_1 = f(M_2 . M_3)$: part correlation

$M_1 . M_2 = f(M_3)$: semipartial correlation

$M_1 . M_2 = f(M_3 . M_2)$: partial correlation

$M_1 = f(M'_2)$: multiple correlation

$M'_1 = f(M'_2)$: canonical correlation

$M'_1 = f(M'_2 . M'_3)$: multiple part correlation

$M'_1 . M'_2 = f(M'_3)$: multiple semipartial correlation

$M'_1 . M'_2 = f(M'_3 . M'_2)$: multiple partial correlation

$M = f(C)$: correlation ratio, one-way analysis of variance

$M_1 . M_2 = f(C)$: one-way analysis of covariance

$M = f(C')$: complex analysis of variance

$M' = f(C)$: multivariate analysis of variance

$C = f(M')$: discriminant analysis

$M'_1 . M'_2 = f(C)$: multivariate analysis of covariance

$M' = f(C')$: complex multivariate analysis of variance

$M'_1 . M'_2 = f(C')$: complex multivariate analysis of covariance

The period means "independent of" in the sense of a least-squares residual. For example, the paradigm for multivariate analysis of covariance can be read, "a set of metric variables, with another set of metric variables partialed out, is expressed as a function of a single categorical variable."

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

NOTE ON STATISTICAL DIFFERENCES

HARL R. DOUGLASS
Dean Emeritus
University of Colorado

UNTIL THE past two decades the reliability of statistical values was determined by the use of the numerical values assigned to properties of the normal probability curve. More recently the reliability of differences, as well as other statistics, is calculated with chi-square and variance ratio. The method, however sound, is very frequently misinterpreted; for example, in many of the research studies the conclusion is that there was no difference between the control and the experimental groups at the 1 percent level, 2 percent level, or 5 percent level, etc. A fairly common interpretation is that there is no difference between the two groups. As a matter of fact, what is usually intended is that there were not sufficient cases to result in an obtained difference beyond the probability of being attributable to chance errors of sampling and measurement in one chance in 100, two chances in 100, or five chances in 100. In other words, the real conclusion in a great many of these studies is that the investigation did not employ enough cases to permit a statistically reliable difference to be concluded.

Furthermore, there is a tendency for writers of research projects to use "reliable difference" in such a way as to give the impression that the size of the difference is not important when, as a matter of fact, what it simply means according to the null hypothesis approach is that one cannot conclude with the small number of cases that there is really a statistically reliable difference. The word *significant* in these studies is very often interpreted to mean *practical* significance with respect to size or importance of the difference. As a matter of fact, all that has been shown is that, with the limited number of cases studied, the difference is not great enough that one may conclude in 95 such studies out of 100, or 98 such studies, or 99 such studies the differences of these sizes would be obtained. There may be no real connection between practical and statistical difference in many studies.

Consider the following illustrations of artificial data:

	Sex	
Score	Male	Female
Pass	20	40
Fail	20	20

Chi-square = 2.78

Retain null hypothesis of no differences at 5 percent level

	Sex	
Score	Male	Female
Pass	40	80
Fail	40	40

Chi-square = 5.51

Reject null hypothesis of no differences at 5 percent level.

Size of chi-square or F is not necessarily a measure of practical significance of an obtained difference.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

ATTITUDES TOWARD SERVICES OF STATE DEPARTMENTS OF EDUCATION[1,2]

W. L. BASHAW, JAMES B. KENNEY,
WILLIAM LANDRUM, and R. ROBERT RENTZ
University of Georgia

and

FOSTER WATKINS

Southeastern Educational Laboratory

ABSTRACT

The structure of attitudes toward services rendered by state departments was studied in six southeastern states with a 70-item scale. Development procedures are described. The scale was administered to randomly chosen samples of superintendents (N = 671), central office personnel (N = 404), principals (N = 627), teachers (N = 3,684), and other local personnel (N = 373). Data from each group were factored and studied for dimensionality. The two factors common to all groups consisted, respectively, of positively and negatively worded items despite the fact that the instrument was developed in a way intended to minimize response set factors. A third factor dealt with university-state relationships.

THIS STUDY was conducted as a part of the evaluation of the Regional Curriculum Project, (RCP), the goal of which is the study of the instructional leadership role of state departments of education in the Southeast. The Regional Curriculum Project is designed as a large, long range field experiment. Twenty-four school systems, four in each of six southeastern states, have been chosen as experimental systems. These systems are receiving intensive consultative, and to some degree, financial assistance by the RCP and their own state department. These 24 systems will be compared eventually with the remaining systems in regard to changes in many criteria.

One of the criteria is attitudes toward state department services. An attitude scale was developed by the authors for the purpose of assessing attitude changes over the experimental period. This paper describes briefly the scale development and the results of the norm testing which is to serve as the pretest attitude data for the RCP. Specific attention is given to the factor analysis results since these results were especially interesting.

SCALE DEVELOPMENT

A large pool of items was compiled in workshops consisting of RCP staff, state department personnel, and public school personnel. The authors edited the item pool to reflect, as nearly as possible, unifactor attitude stimuli.

Items were clustered subjectively into ten clusters reflecting attitude stimuli of particular interest to the RCP project goals. Examples are "Readiness for Change," "Value of Publications," "Relations with Higher Education," and so forth. Within each of the item clusters, items were reworded so that half of the cluster consisted of positively worded items and half were negatively worded items. A pool of 172 items was selected for pilot testing.

After considerable discussion, it was decided to use a standard 5-point Likert scale. The use of a multiple choice format in which scale points were separately defined for each item was a preferred format; however, this format was discarded as being too bulky in terms of test materials. Items were presented as a series of simple, declarative sentences. The response choices were identical for every item and ranged from

> "1--I almost always disagree with this statement; or in almost all instances, this statement is false."

to

> "5--I almost always agree with this statement; or in almost all instances, this statement is true."

Respondents were directed to omit an item only if they had insufficient information upon which to base a response. This definition of omissions was used so that omission patterns could be studied as indicative of areas in which information exchanges were weak.

A pilot sample of 122 graduate students at the University of Georgia was chosen. This pilot sample consisted largely of in-service extension classes and contained representatives of all job categories of interest. Pilot data were analyzed by a series of factor analyses. The 172 items were split into three sets of 57, 57, and 58 items for initial factoring. Items were discarded on the basis of over-complexity (high loadings on several factors), low communalities (implying low reliability as well as identifying specifics), and redundancy. Data were re-analyzed iteratively until a final pool of seventy items was chosen.

The pilot data yielded eight identifiable common factors consisting of five to thirteen items each. Each subset of items was reworded so that half of each subset consisted of positively worded items and half consisted of negatively worded items. The dimensionality of the pilot data was determined largely by nature of the repeated analysis of the same data, and hence is not directly comparable to final results.

SAMPLE

All subjects are public school personnel from six southeastern states. State department records and tape listings were used to identify simple random samples of personnel in four job categories. The percentages of persons in each job category included in the target sample were

1. Superintendents	100 percent
2. Central Office Personnel	12 percent
3. Principals	10 percent
4. Local School Personnel	3 percent

These percentages were higher for the samples drawn from the twenty-four experimental systems.

Test materials were distributed through the six state departments. Subjects returned the materials by mail and stamped envelopes were provided for this purpose.

The percentage return ranged from 55 percent for local personnel to 100 percent for central office personnel. The total percentage return was about 60 percent. Percentages were relatively consistent state-by-state.

The category "local school personnel" was divided into "classroom teachers" and "local personnel who are not teachers" by sorting obtained data.

The sample sizes obtained were as follows:

Superintendents	671
Central Office Personnel	404
Principals	627
Non-teacher, other local personnel	373
Teachers	3,684
Total	5,759

The high percentage return from supervisory persons (near 100 percent in every state) was especially pleasing since this group is the focal point of state-local system relationships.

ANALYSIS

Extensive analysis was made of the data for the RCP; however, this discussion will focus only on the factor analyses. Items were factored by job category with unities on the principal diagonal of the intercorrelation matrices and varimax rotations were used.

The major problem was the determination of dimensionality. Several alternatives have been suggested, the most common being Kaiser's "little jiffy" criterion of using the number of factors corresponding to the number of latent roots exceeding one. Another commonly used criterion is to take the number of factors that are psychologically definable. This non-analytic procedure requires multiple rotational analyses, but is defensible psychologically.

Related to this second criterion is the use of the number of factors that appear consistently over several sets of data, a criterion which is defensible psychologically, but is often useless due to the absence of replication data. This point-of-view has been used to some degree in this study.

The authors chose to follow, as far as possible, Cattell's point-of-view on dimensionality (1: 174-243). Cattell points out (p. 207) that the Kaiser criterion is inadequate for large analyses.

> . . . the Kaiser test. . . cuts off too soon when variables are few ($n < 20$), and cuts off too late when they are many ($n > 50$).

This expectation led to the prediction that our analyses would yield many roots larger than one--an outcome that was obtained and is commonly obtained in item factoring.

Cattell proposes his "scree test" as a semi-analytical solution. His scree test consists of plotting the magnitude of the roots against the number of the root and examining the smaller roots for a straight line trend. He states (p. 206):

> When successive roots extracted begin to fall in this regular way one is dealing only with common factors due to a large number of random small influences. . .

He points out that these "random small influences" would be explained by classic theory as probably specifics orthogonal in the population and measurement error.

He goes on to say (p. 206):

> In large samples, there are usually clear representations of two, not one, successive straight scree slopes. In this case, one takes the line of the upper slope.

TABLE 1

FIRST 25 ROOTS FOR ALL ANALYSES

Root	Group				
	Superintendent	Principal	Other Control	Teachers	Other Local
1	14.49	11.97	12.97	16.41	12.49
2	5.07	9.38	8.76	10.06	10.82
3	3.37	2.68	2.92	1.63	2.20
4	1.88	1.68	1.79	1.41	1.90
5	1.55	1.57	1.57	1.30	1.65
6	1.49	1.41	1.50	1.17	1.47
7	1.39	1.35	1.42	.99	1.46
8	1.32	1.27	1.35	.97	1.39
9	1.28	1.19	1.28	.96	1.34
10	1.24	1.17	1.27	.94	1.23
11	1.18	1.15	1.27	.90	1.22
12	1.12	1.10	1.19	.87	1.17
13	1.10	1.08	1.14	.86	1.10
14	1.07	1.01	1.12	.82	1.06
15	1.03	.99	1.10	.80	1.03
16	.98	.97	1.06	.79	1.01
17	.96	.95	1.01	.76	.97
18	.94	.94	.97	.75	.96
19	.93	.91	.95	.73	.93
20	.91	.89	.91	.71	.89
21	.89	.88	.89	.71	.88
22	.86	.84	.88	.70	.85
23	.85	.80	.86	.70	.84
24	.82	.79	.83	.68	.81
25	.78	.78	.80	.67	.76

RESULTS AND DISCUSSION

A partial table of roots for the five analyses appears as Table 1. The roots from the five analyses were plotted and examined for scree slopes. In each analysis, the lower scree line was obvious. The second scree line appeared clearly in three of the five analyses and less clearly in the other two. Figure 1 shows the typical root graph with two scree lines. Note that in Figure 1 the "root-of-one criterion" intercepts the scree line in a way that would identify a large number of factors, whereas the upper scree line identifies a very small number of factors.

FIGURE 1

TYPICAL SCREE GRAPH SHOWING THREE FACTORS

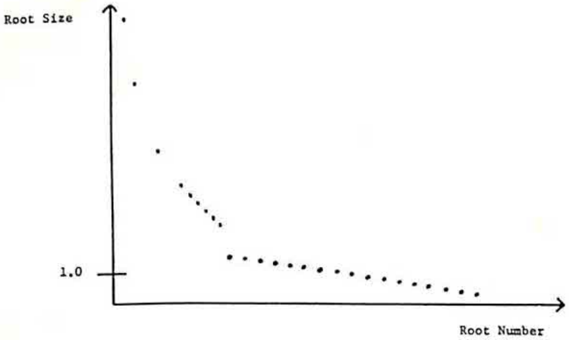


Table 1 shows the number of roots identified by the "greater-than-one" criterion. From 14 to 17 roots exceeded one in four analyses and six roots exceeded one in the teacher data.

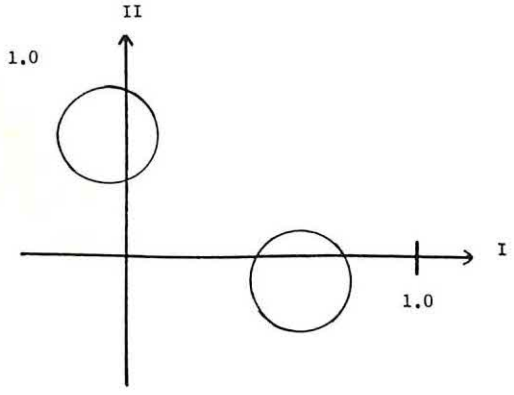
The scree plots were judged as identifying three roots in the data for superintendents, other central office personnel, and principals. Two roots were judged to be important in the data for teachers and other local personnel.

Factor similarity coefficients were calculated for factor matching. Matching factors generally had similarity coefficients in the range .85 to .95.

The factors were examined for identification. At least twenty-five items were found common to all five analyses as identifying both factors I and II. Plots of factor loadings showed that the two factors consisted of mutually exclusive item clusters. A typical plot appears in Figure 2.

FIGURE 2

TYPICAL SCATTERGRAM OF ROTATED LOADINGS ON FACTORS I AND II



The first two factors correlate in the range of 0 to -.3 for four analyses. The correlation is about -.7 for superintendents. The first two factors in all five sets of data define a definite response set splitting positively worded items and negatively worded items into mutually exclusive subsets.

The third factor, identified clearly by six items common to three analyses, consisted of items dealing with the relationship of the state departments to institutions of higher education. The failure of local school people to yield factor III is most likely due to insufficient knowledge at this level to allow consistent responses to these items.

Scales defined by items loading highly on each factor were analyzed for internal consistency reliability by calculating alpha coefficients for thirty job-by-state subgroups. Alpha coefficients for factors I and II were in the range .84 to .95, 25 of the thirty coefficients exceeding .90 for both factors. Factor III alpha coefficients were lower, being largely in the range .70 to .87. Test-retest reliabilities over a 1-month interval were obtained on a sample of seventy-two and were all in the mid-seventies.

Factors I and II clearly indicate that subjects responded independently to positive and negative items. The authors have been unable to determine any artifact that would force this solution. Skewness in the distributions of item responses would not create the split, since the skewness has the effect of forcing pear-shaped joint distributions on all item pairs, thus affecting correlations in a fairly homogeneous way. The skewness certainly can affect the total structure by lowering correlations of item pairs that correlate negatively in the population when both items are either positively or negatively worded. However, this possibility does not seem to be related to the result of independence between positive and negative items.

This result has also been obtained in other studies. Some of the authors of this paper found such an item split in two of several factors identified in an independent study of the attitudes of teachers toward their county system services. In a personal communication to the senior author, Garry Foster (2) of the Nova School Project reports getting several factors in children's responses to items about their attitudes toward education. Two of Foster's factors were negative and positive attitudes.

A suggestion of this response set is also seen in Kerlinger's (3) study of progressive and traditional attitudes toward education. Kerlinger reports two independent major attitudes as second order factors--namely, progressive attitudes and traditional attitudes. A review of the items loading on these two factors suggests that the response set found by the present authors could have been operating in Kerlinger's study. Kerlinger's items could be identified as negatively worded or neutrally worded items. There were no positively worded items. In general, the negatively worded items loaded highly on only the "traditional attitude" factor.

Whether or not the positive-negative response set is a methodological problem to be overcome or a substantive finding remains to be seen. Certainly one can avoid the set by constructing items having only positively, negatively, or neutrally worded items without mixing the item types. It can also be avoided by choosing alternate item formats such as a semantic-differential-like item format. However, such scales might serve merely to hide what is perhaps a significant and substantive effect--a tendency to evaluate independently the good and the bad aspects of an educational stimulus.

FOOTNOTES

1. Supported by the USOE as part of the evaluation of the Regional Curriculum Project, 50 Whitehall Street, Atlanta, Georgia, directed by Edward T. Brown. The conduct of the study and opinions expressed are solely those of the authors and do not reflect official opinions of the RCP.

2. Authors are listed alphabetically. Appreciation is expressed for the assistance of Edward T. Brown and Edith Miller, Regional Curriculum Project, H. T. Conner, Richardson Foundation, and the many persons in each of the six states who assisted in the data collection.

REFERENCES

1. Cattell, Raymond B., "The Meaning and Strategic Use of Factor Analysis," in Cattell, Raymond B. (Ed.) Handbook of Multivariate Experimental Psychology, Rand-McNally, Chicago, 1968.

2. Foster, Garry, personal communication, December 18, 1967.

3. Kerlinger, Fred N., "The First/and Second-order Factor Structures of Attitudes Toward Education," American Educational Research Journal, 4: 191-206, May, 1967.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

THE RELATIONSHIP BETWEEN SELF CONCEPT AND ACADEMIC ACHIEVEMENT

MORRIS D. CAPLIN
Queens College

ABSTRACT

It was hypothesized that children, both white and Negro, attending a de facto segregated school have less positive self concepts than do children attending desegregated schools, and that there is a significant positive relationship between self concept and academic achievement. Sixty children from the intermediate grades of each of the elementary schools in a small city in northern New Jersey were matched on the basis of age, grade, sex, race, intelligence, and socio-economic status. Analyses of variance were computed on the scores obtained from the self-report instrument administered and correlations between these scores and achievement scores were calculated. It was found that children attending the de facto segregated school had less positive self concepts. There was also a significant positive relationship between self concept and academic achievement. That is, those children having more positive self concepts had higher academic achievement.

THE SUPREME COURT decision of 1954 outlawing de jure school segregation recognized that schools cannot be separate and equal because this condition contributes to psychological degradation and formation of a negative self-image in Negro children who are rejected and isolated in racially segregated schools. In recent years the meaning of desegregation has been applied also to racial imbalance not due to legal discrimination. Many recent court decisions have found that a predominance of minority-group children constitutes de facto segregation.

Various educational trends have evolved to meet the changing needs of our society. The compulsory education law was enacted to eliminate illiteracy and industrialization gave rise to vocational schools. The educator thus was faced with continuously adjusting educational approaches to meet the challenge of a changing society. Social change, both national and international, has placed a new demand on education - providing for the culturally disadvantaged, particularly the Negro. The first and most important step may be desegregation of the de facto segregated schools.

De facto segregated schools in some substantial measure characterize most public schools in northern cities (6). A study was conducted to determine whether children attending such a school have less positive self concepts than children attending desegregated schools, and whether there is a significant positive relationship between self concept and academic achievement (5).

INTRODUCTION

That personality factors may exert a powerful influence on academic achievement has long been recognized. Harris (8) compiled a list of over 300 studies concerned with this topic in 1940. Factors most frequently mentioned as important for academic success were ambition, emotional security, a sense of responsibility, cooperativeness, and seriousness. More recently, however, the significance of the self concept has been stressed. The importance of the self concept in relation to academic learning has been given particular emphasis by Lecky (11). Substantial evidence indicates a relationship between self concept and academic achievement. The literature shows that most of the studies have been conducted at the high school or college level; the elementary school has seen relatively few investigations of this problem.

It would seem that the relation between self concept and academic achievement should be quite high and positive. Those with high achievement know it, and those with low achievement know this too. Therefore, those who achieve well should have the confidence which results from the practice of succeeding and those who do not do so well do not have as much self confidence and may not rank themselves so highly. The question arises as to what the precedent variable was. That is, does one achieve higher because of a higher self concept or does one have a higher self concept because he achieved higher? Jersild (9) expresses the point of view that the learner proceeds, interprets, accepts, resists, or rejects what he meets in school in the light of a self system he has within him.

In studying the relationship of self concept to achievement in reading of fifth-grade children, Lumpkin (12) found significant relationships between pupils' self concepts and achievement in reading.

As measured by the instruments employed, overachievers revealed significantly more positive self concepts. In his conclusion Lumpkin states that this outcome confirms current opinion that achievement stems from intrinsic motivation, as well as from the environmental responses to this achievement.

Based on the reported association between poor self concepts and reading disabilities, Wattenberg and Clifford (15) did an exploratory study in an effort to determine which was the antecedent phenomenon. One of the hypotheses states that the self concept is more important as a determinant of learning efficiency than learning experiences are in the formation of self concept. They concluded that in general, the measures of self concept and ego strength taken at the kindergarten level were predictive of reading achievement two and one-half years later and that even as early as kindergarten, self concept phenomena are antecedent to and predictive of reading accomplishment at the least.

Academic grade-point averages of children in grades 3 through 11 were found by Bruck (8) to be influenced by self concept. He concluded that a positive and significant relationship exists between self concept and grade-point average on all grade levels ranging from 1 to the 5 percent level of significance.

Bodwin (9) did a study on the relationship between self concept and certain educational disabilities of children in grades 3 through 6. He found that a positive and very significant relationship existed between immature self concept and reading disability, and a positive and very significant relationship existed between immature self concept and arithmetic disability.

In other related studies, Kurtz and Swenson (10), Nason (13), and Blackham (1) found that underachievers had lower self concepts than did achievers.

INSTRUMENT

The self concept has been a concern of psychology for a number of years. Despite this, the subject remains elusive and confusing, problably because of the lack of precision in defining this concept. The lack of a really acceptable definition of self concept contributes to the lack of a really effective measurement technique. Since the self concept is an organization within the individual's perceptual or phenomenal field, it is not open to direct observation. To study the self concept it is necessary to infer its nature from observations of the behavior of the individual. One class of behaviors which may be used as a basis for making inferences, of course, is what the subject has to say about himself.

It seems likely that self-report techniques can provide some insight in the area of self concept and this is the type of instrument used in this study. Its present form was developed at the Horace-Mann-Lincoln Institute of School Experimentation, Teachers College, Columbia University, under the guidance of Professor Miriam Goldberg. Fifty items were selected which seemed to present clear examples of self concept definitions. The items were derived, in part, from a study of the self-descriptions of intermediate grade pupils, at which time each trait was designated as positive or negative on the basis of inter-judge agreement (7). A number of intermediate grade school children from various New York City schools were used as judges. Other items, especially those describing pupil abilities were included on the basis of teacher experience. The intermediate grade teachers were also from various New York City schools and also judged the items for appropriate reading and conceptual levels.

Sixty percent of the items concern the pupil's feelings about himself as a person and/or social being. Forty percent relate to his feelings about school. Classification of items as personal/social or school-related was made by a panel of intermediate-grade teachers from the school system in which this study was made.

SAMPLE

The sample group consisted of 180 intermediate grade children from the three elementary schools of a small city in northern New Jersey located near New York City. One of these schools was de facto segregated with a Negro enrollment of over 66 percent. The second school was newly desegregated by transferring 150 Negro children to this previously almost all-white school. The third school had been desegregated for many years due to the housing pattern of that neighborhood.

Since the long-term desegregated school had the smallest Negro enrollment (36 intermediate-grade pupils) of the three schools, this group served as the base group. Thirty out of the 36 Negro pupils were selected for the base group. The total group of 36 was not used because it was not possible to match six of these pupils on all of the control variables to pupils in the other schools. This base group consisted of 15 Negro girls and 15 Negro boys, 10 in the fourth grade, 13 in the fifth grade, and 7 in the sixth grade. A similar group of white pupils from the long-term desegregated school was matched to the base group on the basis of grade, sex, intelligence, and socio-economic status.

Sixty children from the de facto segregated school, 30 white and 30 Negro, were then matched to the base group again on the basis of grade, sex, intelligence, and socio-economic status. Finally, an equal number, 30 white and 30 Negro from the newly desegregated school, was matched the same way.

PROCEDURE

Analyses of variance were computed on the scores representing self concept for the entire group, for the girls separately, for the boys separately, among the white pupils, among the Negro pupils, between the white and Negro pupils, and between the girls and boys. The entire instrument was then subdivided into two categories and analyses of variance were computed as above for those items that centered primarily on the pupil's feelings about himself as a personal and/or social being, and for those items that related directly to his feeling about school.

Correlations between the scores on the self concept instrument and the standard composite scores on the Iowa Test of Basic Skills were calculated. Correlations were also computed for the personal/social qualities and for the school-rated items.

FINDINGS

It was found that the school-rated self concepts of children, white and Negro, attending the de facto segregated school were significantly lower than those of children attending the newly-desegregated or long-term desegregated schools. It was also found that children having more positive self concepts had higher academic achievement. The correlations obtained between the composite standard scores on the Iowa Test of Basic Skills and the self concept scores on the entire instrument was . 52; for personal/social qualities . 45; and for school-related items . 58. All three correlations were significant at or beyond the . 001 level of significance. These findings lend strong support to the significant positive relationship between these measures. From this finding it could not be said that academic achievement is determined by the self concept. However, while a correlation does not imply a cause-effect relationship, it does at least give educators and psychologists the right to look for a cause-effect between self concept and academic achievement.

In considering the two relatively distinctive aspects of self concept measured in this study, the question arises: In relation to academic achievement, are the most significant aspects the generalized feelings and attitudes about the self, as reflected by the personal/social scores, or are specific self perceptions, as reflected by the school-related scores, more significant? Apparently school-related traits are more intimately related to academic achievement than are the generalized personal/social feelings about the self.

DISCUSSION

When the significant differences between schools are compared, there appears to be less difference between the de facto segregated and the long-term desegregated schools than between the de facto segregated and the newly-desegregated schools. That is, the children attending the newly-desegregated school were significantly higher on school-related self concept. This interpretation is based upon the assumption that the children in the newly-desegregated school did not have a significantly higher school-related self concept to begin with. This assumption seems to be reasonable since the Negro and white children who left the de facto segregated school were transferred solely on the basis of a shift in boundary lines. Thus, all Negro and some of the white pupils in the sample group from the newly-desegregated school had attended the de facto segregated school previously.

This finding suggests that a certain process may occur when children are transferred to predominately white schools in order to achieve racial balance. First, the school-related self concept appears to go up. This might be due to the initial effect, a type of Hawthorne effect, due to the novelty of being in a different situation and different school climate. Then, with the passage of time, the Hawthorne effect may decrease and school-related self concept goes down.

What may occur is that the child finds himself in a desegregated school which may symbolize for him an opportunity. His school-related self concept goes up. Now the child sees his inadequacies magnified in the heterogeneous classroom where most of the pupils are obviously better achievers. He finds that his achievement does not increase, and in fact, may decrease as is often the case with low achievers. Once a pattern of low achievement has been established, it generally continues and becomes worse. This could cause his school-related self concept to drop again, perhaps even more severly. Thus, in this study involving a specific elementary school population, the apparent effect of desegregation was an initial upsurge in school-related self concept. The school must take positive action to protect and enhance this initial gain. Even though there is no evidence to show a direct casual relation one to the other, it seems likely that a higher self concept may contribute to higher achievement and higher achievement in turn may contribute to a higher self concept.

ADDITIONAL FINDINGS

Two additional analyses of variance were computed with the data available. While these are not offered in support of the hypotheses, they are of interest and may provide a base for further research in the area. First was an examination of the differences in self concept between the 90 girls and 90 boys in the sample group. No significant differences were obtained on any of these analyses.

This finding may seem contrary to general expectations since it is well established that girls of ages 9-11 are physically and physiologically more developed than boys. Also, more frequent contacts with women teachers, who predominate in the elementary school, and with their mothers may have enabled girls to develop a more satisfying self concept at these developmental stages.

The literature on sex differences is vast, but relatively few studies have been made in the area of self concept in the elementary school. Some previous studies have indicated that girls may perceive themselves as significantly more accepted and intrinsically valued than boys (2). On the other hand, Bruck, in his study of sex differences in the relationship between self concept and grade-point average in grades 3-9 concluded that there was no significant sex difference in this relationship when one compares elementary pupils as a group (4).

In a study by Piers and Harris, a wide-range self concept instrument was developed and administered to four third-grade classes, four sixth-grade classes, and four tenth-grade classes. In the elementary grades, classes were chosen from several different schools, representing a cross-section of socio-economic levels in the community. Sex was found not to be a significant variable (14).

Wylie points out that in many of the studies, as in the study reported here, exploring sex differences was not the main purpose, and it is probable that dependent variable measures were not specifically chosen for their relevance to a study of sex role (16).

The second additional analysis was an examination of the differences in self concept between the 90 white and 90 Negro pupils in the sample group. Significant differences were indicated with the Negro group being lower at the .05 level of significance.

Wylie (17), in a study concerning associations between self-evaluation and several variables, found significant differences between white and Negro groups. However, these were not matched groups and Wylie goes on to state, "The significant differences which were obtained may plausibly be a function of the occupational-level differences between the total Negro and white groups."

The significant differences obtained in this study were probably not a function of occupational-level differences since the white and Negro pupils were matched on socio-economic level. However, when the effects of school and race upon self concept were examined together, the significant difference was accounted for in large measure by the Negro children attending the de facto segregated school. No definite conclusions can be drawn because of the lack of clearly positive findings.

SUMMARY

The major focus of this study was academic achievement and its relation to self concept. It was hypothesized that children, both white and Negro, attending a de facto segregated school have less positive self concepts than do children attending desegregated schools; and that there is a significant positive relationship between self concept and academic achievement. Analysis of the data supported these hypotheses.

REFERENCES

(1) Blackham, Garth J. "A Clinical Study of the Personality Structures and Adjustment of Pupils Underachievement and Overachievement in Reading," PhD Dissertation, Cornell University, 1954, 167 pp. Dissertation Abstracts, XV (1955), 1199 pp.

(2) Bledsoe, Joseph C. "Sex Differences in Mental Health Analysis Scores of Elementary Pupils," Journal of Consulting Psychology, XXV (July, 1961), pp. 364-365.

(3) Bodwin, Raymond F. "The Relationship Between Immature Self Concept and Certain Educational Disabilities," PhD dissertation, Michigan State University, 1957, 67 pp. Dissertation Abstracts, XIX (1959), 1645 pp.

(4) Bruck, Max. "A Study of Age Differences and Sex Differences Between Self Concept and Grade-Point Average," PhD. dissertation, Michigan State University, 1957, 98 pp. Dissertation Abstracts, XIX, (1959), 1646 pp.

(5) Caplin, Morris D. "The Relationship Between Self Concept and Academic Achievement and Between Level of Aspiration and Academic Achievement," EdD dissertation, Teachers College, Columbia University, June, 1966.

(6) Commission on School Integration. Public School Desegregation and Integration in the North, published as a special issue of the Journal of Intergroups Relations (Washington, D. C., November, 1963), pp. 13-21.

(7) Davidson, Helen and Balducci, Don. Check List of Trait Names, unpublished paper, City College of New York, 1955.

(8) Harris, David. "Factors Affecting College Grades: A Review of the Literature, 1930-1937," Psychology Bulletin, XXXVII (March 1960), pp. 151-166.

(9) Jersild, Arthur. Child Psychology (New York, Prentice-Hall, 1954), p. 14.

(10) Kurtz, John J. and Swenson, Esther J. "Factors Related to Over-achievement in School," School Review, LIX (November, 1951) pp. 472-480.

(11) Lecky, Prescott. Self Consistency: A Theory of Personality (New York, Island Press, 1954).

(12) Lumpkin, Donovan D. "The Relationship of Self Concept to Achievement in Reading," EdD, dissertation, University of California, 1959, 175 pp. Dissertation Abstracts, XX, (1959), 214 pp.

(13) Nason, Leslie J. Academic Achievement of Gifted High School Students (Los Angeles, University of Southern California Press, 1948), 92 pp.

(14) Piers, Ellen V. and Harris, Dale B. "Age and Other Correlates of Self Concept in Children," Journal of Educational Psychology, LV (April, 1964), pp. 91-95.

(15) Wattenberg, William W. and Clifford, Clare. "Relations of Self-Concepts to Beginning Achievement in Reading," Child Development, XXXV, (June, 1964), pp. 466-467.

(16) Wylie, Ruth C. The Self Concept (Nebraska, University of Nebraska Press, 1961), 145 pp.

(17) Wylie, Ruth C. "Children's Estimates of Their Schoolwork Ability as a Function of Sex, Race, and Socio-economic Level," Journal of Personality, XXXI, (June, 1963) 206 pp.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

THE CREATION AND SOLUTION OF UNIQUE STATISTICAL PROBLEMS FOR EACH STUDENT

GEORGE COUNTS

Cape Girardeau, Missouri

ABSTRACT

This study was designed to answer these questions: Can useful computer programs be developed? If such programs can be developed, how much storage is needed and what is the cost per student? Two Fortran programs were developed that created and analyzed data. The two types of analyses were correlational analysis and a *t*-test of differences between means for independent samples. Each program generated and solved thirty problems. A sample problem and the associated statistical analysis were presented in tabular form for each program. The computed values of *r* (and *t*) for each student were reported. These techniques could be used for other types of problems without exceeding the storage of many computers. The cost was substantial.

AS THE BURDEN on the school becomes more complex, methods of instruction must be improved. Many writers (1, 3, 4) would make the situation manageable by tailoring methods and materials to the individual learner. Others (2) place emphasis upon reducing the complexity of materials to fit a large group of students. A slightly different approach is research on the effect of unique materials with identical levels of complexity. The effect of such materials may improve the mastery of content; and the behaviors exhibited during subsequent learning experiences or during normal professional activities may relate to course objectives.

In an introductory course on statistical methods in education and psychology appropriate materials for individualizing instruction have been difficult to prepare. If a single textbook were adopted, it normally included at most one or two problems that illustrate a particular technique. To lift examples from other texts was and is ethically questionable, and to develop sets of new but similar data time consuming. The professor who generated such data then faced the problem of computation. Even with access to the speed of the electronic computer, the individual who prepared the input data had a long tedious chore. A final consideration has been the willingness of students to prepare files of problems. As a result, materials developed and used in the past must be assumed to be accessible for at least some of the students.

The interface between the teaching of courses in statistics and utilization of the computer is an area not fully explored. There is a danger that we may use the computer as if it were only a fast desk calculator. The computer has been used extensively in data analysis. We may also use it to generate random numbers with a normal or rectangular distribution. However, greater integration should be possible. Random numbers generated as output are subject to analysis as raw data. If storage permits, the best use of computer time is to complete the analysis during the same run as the data were generated. The speed of the computer may permit such a program to prepare unique (but similar) problems for each class member at a reasonable cost. These problems would be unique in the sense that the raw scores are probably different but each set of scores could constitute a sample from the same universe. Even if research had clearly defined the most effective alternatives for the era before the advent of the electronic computer, the choice of alternatives following its development would still be fluid.

STATEMENT OF THE PROBLEM

As the theoretical justification for the use of materials is irrelevant unless the materials can actually be produced, the scope of this study was restricted to defining statistical relationships, to computer program development, to calculation of core storage requirement, and to cost analyses. The specific questions for which answers were sought were:

1. Can useful computer programs be developed? (The steps in such a program would be to generate unique raw data; to read the student's name; to print the student's name; to print the data in a sensible format on the same page; to analyze the data; to change pages; to print the student's name, the results of the analysis, and the population parameters; to change pages;

TABLE 1

DATA GENERATED BY PROGRAM I FOR JOE JAMES

	JOE JAMES	
N	X	Y
1	45.	55.
2	41.	68.
3	61.	99.
4	35.	69.
5	51.	113.
6	40.	66.
7	27.	58.
8	50.	63.
9	40.	51.
10	63.	63.
11	51.	27.
12	43.	78.
13	49.	71.
14	62.	67.
15	57.	78.
16	47.	53.
17	48.	74.
18	43.	80.
19	54.	31.
20	54.	68.
21	34.	51.
22	49.	48.
23	47.	46.
24	46.	61.
25	52.	82.
26	57.	74.
27	57.	83.
28	37.	41.
29	38.	65.
30	42.	60.
31	49.	91.
32	71.	96.

and to repeat the process for each student in in the class.)

2. If such programs can be developed, what storage requirements are necessary?

3. If such programs can be developed, what cost per student is required for specified types of problems?

PROCEDURES

The IBM 360 model computer of Southeast Missouri State College was used to run the trial programs. The GAUSS and RANDU subroutines listed in System/360 Scientific Subroutine Package Programmer's Manual were provided by the computer center. From these two subroutines, parallel subroutines were written with different storage locations. These subroutines were given the names GAUS and RAND. (These four subroutines are specific to the IBM 360.) RANDU and RAND were random number generators with rectangular distributions, but GAUSS and GAUS created random numbers with normal distributions.

Two main programs were developed. In one program (Program I) for each X_i, Y_i was functionally related to X_i but allowed to vary within specific limits. (These limits were fixed for a particular run, but could be changed for the next pass.) The data created by this technique were subjected to correlational analysis by the same program. In the second program (Program II) population means and standard deviations were allowed to vary within prescribed limits. The total sample was drawn for group X before population values for group Y were created internally. (With a small modification of the program, the number of cases in the sample could also be computed internally but allowed to vary within defined limits.) This program created data appropriate for a t-test of differences between two means for independent samples and completed the analysis. The raw data and the analysis were printed under the control of the instructions in the program.

TABLE 2

ANALYSIS OF RAW SCORES GENERATED BY PROGRAM I FOR JOE JAMES

```
                JOE JAMES

        SUM OF X              1540.00
        SUM OF Y              2130.00
        SUM OF X.Y          104533.00
        SUM OF X.X           76892.00
        SUM OF Y.Y          152770.00

        MEAN FOR X=             48.13
        MEAN FOR Y=             66.56

       (PEARSON PRODUCT-MOMENT)
       CORRELATION = 0.367

     EQUATION FOR PREDICTING Y FROM X
          Y'=          0.729 X +  31.471

     EQUATION FOR PREDICTING X FROM Y
          X'=          0.184 Y +  35.852

     STANDARD DEVIATIONS
            X                     Y
          9.47                  18.83

     STANDARD ERRORS
            Y'                    X'
         17.52                   8.81
```

To determine the storage requirements of each program the amount of core required for both the main program and the subroutines was recorded from the Fortran compiler reports.

In estimating the probable cost per student for computer time, each program was timed during execution. Thirty student name cards were prepared and each of the programs was used to prepare a unique problem for each student. Each program was timed from the reading of the first program card to the printing of the last analysis. The total cost was determined by multiplying the total time in minutes by one dollar per minute. The cost per student is the result of dividing the total cost by thirty.

RESULTS

To illustrate the type of computer output generated by the two programs, representative examples are presented in Tables 1 - 7. Table 1 is a sample problem for correlated data. A fictitious student's name illustrates the means of identification. In Table 2 the summary of the correlational analysis is for the data in Table 1. Table 3 contains actual predictions based upon the summary statistics in Table 2. Table 4 contains the various Pearson product-moment correlation coefficients generated for thirty "students." In each of the problems in the set $\mu_x = 50$, $\sigma_x = 10$, $\mu_y = 70$ and $Y_i = X_i + a_i$. a_i

TABLE 3

PREDICTED RAW SCORES FOR DATA IN TABLE 1

N	X	Y	X'	Y'
1	45.00	55.00	45.99	64.28
2	41.00	68.00	48.39	61.37
3	61.00	99.00	54.11	75.95
4	35.00	69.00	48.57	56.99
5	51.00	113.00	56.69	68.66
6	40.00	66.00	48.02	60.64
7	27.00	58.00	46.55	51.16
8	50.00	63.00	47.47	67.93
9	40.00	51.00	45.26	60.64
10	63.00	63.00	47.47	77.41
11	51.00	27.00	40.83	68.66
12	43.00	78.00	50.23	62.83
13	49.00	71.00	48.94	67.20
14	62.00	67.00	48.21	76.68
15	57.00	78.00	50.23	73.03
16	47.00	53.00	45.62	65.74
17	48.00	74.00	49.50	66.47
18	43.00	80.00	50.60	62.83
19	54.00	31.00	41.57	70.85
20	54.00	68.00	48.39	70.85
21	34.00	51.00	45.26	56.26
22	49.00	48.00	44.70	67.20
23	47.00	46.00	44.33	65.74
24	46.00	61.00	47.10	65.01
25	52.00	82.00	50.97	69.39
26	57.00	74.00	49.50	73.03
27	57.00	83.00	51.16	73.03
28	37.00	41.00	43.41	58.45
29	38.00	65.00	47.84	59.18
30	42.00	60.00	46.91	62.10
31	49.00	91.00	52.63	67.20
32	71.00	96.00	53.55	83.24

TABLE 4

CORRELATION COEFFICIENTS FROM PROGRAM I FOR THIRTY STUDENTS

N	r	N	r
1	.378	16	.489
2	.326	17	.547
3	.641	18	.175
4	.536	19	.367
5	.421	20	.285
6	.532	21	.547
7	.362	22	.568
8	.570	23	.354
9	.616	24	.398
10	.356	25	.129
11	.403	26	.484
12	.514	27	.242
13	.490	28	.160
14	.511	29	.604
15	.384	30	.514

TABLE 5

RAW SCORES GENERATED BY PROGRAM II FOR JAMES JAY

	JAMES JAY	
	X	Y
1	11.00	47.00
2	32.00	55.00
3	30.00	41.00
4	19.00	46.00
5	15.00	49.00
6	25.00	51.00
7	21.00	41.00
8	25.00	47.00
9	22.00	47.00
10	18.00	43.00
11	16.00	54.00
12	19.00	48.00
13	21.00	44.00
14	27.00	47.00
15	24.00	45.00
16	31.00	49.00
17	43.00	46.00
18	26.00	50.00
19	19.00	44.00
20	10.00	45.00
21	32.00	43.00
22	11.00	47.00
23	42.00	42.00
24	31.00	51.00
25	21.00	49.00
26	23.00	44.00
27	21.00	42.00
28	18.00	49.00
29	48.00	43.00
30	21.00	53.00

was a random number with a mean of 20 and a standard deviation of 20. That is, $\sigma_{y.x} = 20$. In Table 5, a representative example of data generated by the second program is presented. Table 6 contains the results of the analysis and population parameters for the data in Table 5. In this program μ_x and μ_y was allowed to randomly vary between 20 and 59, and σ_x and σ_y between 2 and 11. In Table 7 the individual *t* values may be examined. Given a critical value of *t* = 2.00, 26 of the 30 values would allow rejection of an appropriate null hypothesis. If fewer significant values were desired, a reduction in the range of means would give this result.

TABLE 6

ANALYSIS OF DATA GENERATED BY PROGRAM II FOR JAMES JAY

JAMES JAY		
MEAN FOR X	24.07	
MEAN FOR Y	46.73	
STANDARD DEVIATION - X		9.12
STANDARD DEVIATION - Y		3.77
T	-6.60	
PARAMETER	X	Y
MEAN	22	47
S	10	5

The program that generated correlated data required substantially more core storage than the other program. The total required for storage of Program I was 20,904 bytes. Program II required 13,687 bytes. These totals are exclusive of core storage required for the supervisor program.

The total execution time for Program I was 16 minutes, 13 seconds and for Program II was 9 minutes, 44 seconds. The total cost for Program I was $16.25 and for Program II was $9.75. The total cost per student for a correlation problem was 54¢, and the total cost per student for the *t*-test problem was 33¢. Additional refinement of procedures can slightly reduce the cost per student, but the operating cost of a program should not be confused with the cost of development.

CONCLUSIONS

Two functional programs were developed. By modifying the statistical relationships, by changing the shape of distribution, and by varying statistical analyses, many types of problems can be given similar treatment. The amount of storage required for comprehensive programs is not beyond the capacity of many types of computers. The cost per student is substantial; however, if at any time the computer stands idle, the real cost may just be the cost of operation.

TABLE 7

COMPUTED *t* VALUES FROM PROGRAM II FOR THIRTY STUDENTS

N	*t*	N	*t*
1	6.15	16	-1.94
2	7.04	17	-1.87
3	-6.60	18	5.38
4	6.62	19	-2.49
5	-7.17	20	5.95
6	-.28	21	3.21
7	-7.25	22	5.67
8	7.44	23	-6.41
9	-7.07	24	4.35
10	3.51	25	7.67
11	6.93	26	-7.29
12	5.92	27	3.59
13	-6.61	28	5.77
14	5.79	29	6.89
15	-6.95	30	-1.36

It appears that these problems are different from most textbook problems. It is possible that problems with known parameters will help to clarify the rationale of sampling from populations with unknown parameters. The population values are accessible to the professor, and he can control the timing of the access by students.

Travers (5:78) implies that one who develops materials should not attempt to evaluate such materials. To facilitate research neither of the two programs are being placed under copyright, and the author welcomes correspondence on this subject.

REFERENCES

1. Bundy, Robert F., "Computer-Assisted Instruction - Where Are We?" *Phi Delta Kappan*, 49:424-429, April 1968.
2. DeHaan, Robert F., and Doll, Ronald C., "Individualization and Human Potential," *Individualizing Instruction*, pp. 9-28, edited by Ronald C. Doll, 1964 Yearbook of the Association for Supervision and Curriculum Development, N.E.A., the Association, Washington, 1964.
3. Rath, Gustave J., "Non-CAI Instruction Using Computers and Non-Instructional Uses of CAI Computers," *Educational Technology*, 8:11-13, February 1968.
4. Suppes, Patrick, "The Uses of Computers in Education," *Scientific American*, 215:206-208, 213-220, September 1966.
5. Travers, Robert M. W., *An Introduction to Educational Research*, The Macmillian Co., New York, 1964.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

TEACHERS' RATINGS, SEX, AND SCAT AS PREDICTORS OF RANK IN HIGH SCHOOL CLASS[1]

DR. PAUL W. DIXON, NOBUKO K. FUKUDA
University of Hawaii Hilo Campus
and
ANNE E. BERENS
University of Hawaii

ABSTRACT

Data from 561 high-school seniors were analyzed using multiple linear regression to determine the independent contribution of each variable to the criterion of rank in class (R). Significant F ratioswere obtained for SCAT V and Q, teachers' ratings of Accuracy (ACCU), Effort and Industry (E-I), and Reliability and Responsibility (R-R). R-R and ACCU, but not E-I, were found to be significant preditors of SCAT V and SCAT Q. Females were significantly higher than males on R, ACCU, E-I, and R-R and lower on SCAT Q. The data were interpreted as supporting the effectiveness of teachers' ratings in predicting academic performance in high school. Females' superior performance on teacher ratings was interpreted as evidence that females conformed more than males to teachers' expectations.

AUSUBEL and others (1), found significant positive correlations between teachers' ratings of adjustment, persistence, scholastic competitiveness, academic aspirations, and the criterion of grade point averages in junior high school. The correlations were significant for both boys and girls with neither boys nor girls showing higher correlations in these attributes.

According to the manual (Cooperative Test Division, 1955, pp. 3-5) the School and College Abilities Tests (SCAT) are designed for the purpose of helping teachers and counselors estimate the capacity of individual students to perform work at the next level of schooling.

The items of the test are designed to measure school-learned aptitudes rather than general intelligence. The reason given for this is that the best single predictor of success from one year to the next is the performance observed in the preceding year. The manual also states that verbal and quantitative skills appear to be the critical prerequisites to next steps in learning in general education.

Williamson and Cole (8) report that indices on the Behavioral Differential, following the format of the semantic differential (5), predict academic success better than measures of hypothetical entities, such as intelligence. Their findings suggest that teachers' ratings may have predictive validity if the scales are correctly formulated and are bipolar in nature.

This report is concerned with: (a) Whether a school-generated teacher rating scale predicts rank in high-school class; (b) Whether SCAT scores predict rank in class; (c) Whether SCAT and Teachers' Ratings differ in their predictive validity; (d) Whether membership in sex grouping significantly affects SCAT scores, rank in class, or teachers' ratings. Findings which support the predictive validity of teachers' ratings suggest that additional weighting be given to this type or measurement in a high-school setting.

Subjects

The subjects were members of the graduating class of a large high school in Hawaii. The students who were selected for study had complete records on measures of SCAT, teachers' ratings, and rank in class. The ethnic groups retained for analysis were the Japanese, Filipinos, Portuguese, and Cosmopolitans (one person representing two or more ethnic groups). Ethnic groups consisting of less than ten per group, such as Chinese, Koreans, Causasians, and Hawaiians, were not used in the analysis. From the graduating class of 643 students, 561 were used in this study.

TABLE 1

TEACHERS' RATING SCALE

	0 10 20	30 40 50	60 70 80	90 100
ACCURACY (ACCU)	Work is poor. Makes frequent errors.	Work is inaccurate and below standard.	Work is well done and reasonably accurate.	Work is of highest quality.
COOPERATION (COOP)	Disagreeable. Cannot or will not work with others.	Works with others sometimes but has difficulty.	Usually agreeable. Generally willing to help.	Always agreeable. Willing to do extra favors.
EFFORT-INDUSTRY (E-I)	Does as little as possible. Lazy.	Seldom completes required work.	Usually does work that is required. Occasionally does extra work.	Very industrious. Does extra work gladly.
INITIATIVE-LEADERSHIP (I-L)	Acts only under direction.	Seldom originates any work. Follows others.	Plans many of his activities and executes them. Still needs supervision.	Marked ability to think for himself. Carries out ideas.
RELIABILITY-RESPONSIBILITY (R-R)	Neglects promise and obligations. Unreliable.	Reliable on some occasions. Often needs supervision.	Usually dependable. Conscientious.	Thoroughly dependable.
PROMPTNESS-PUNCTUALITY (P-P)	Undependable. Almost always late.	Frequently late.	Usually on time but occasionally late.	Always on time.
SELF-CONFIDENCE (S-F)	Timid. Hesitant. Easily influenced.	Appears to be over self-conscious.	Wholesomely self-confident.	Shows superb self-assurance.

Instrumentation

The records for the graduating class were examined and SCAT scores, rank in class, and teachers' ratings were recorded. These records are compiled during the first semester of the senior year by the school principally for college transcript requests. The typical student had been rated by five teachers, resulting in five ratings per scale with the range being three ratings to six ratings (see Table 1). The rating system thus conformed to the criteria given by Rugg (6) who recommends the use of at least three judges and approaches the number set by Symonds (7) who recommends the use of eight judges. The ratings made by the teachers were not confined to the intervals of the scale such as 0, 10, or 20 but were placed at any point which the teacher considered appropriate.

Procedure

The marks placed by the teachers on the rating scale of each student were measured from the zero point on the scale in centimeters to one place accuracy. Thus a student might have scores of 9.6, 9.6, 9.8, 8.5 and 8.9 as measured in centimeters. For each student, ten different scores were obtained: Verbal (SCAT V), Quantitative (SCAT Q), averages of the teachers' ratings on the scales of Accuracy (ACCU), Cooperation (COOP), Effort and Industry (E-I), Initiative and Leadership (I-L), Reliability and Responsibility (R-R), Promptness and Punctuality (P-P), and Self-Confidence (S-F), and the rank in class (R).

RESULTS

The subroutines (PERSUB) of Bottenberg and Ward (2) and an IBM 7040 computer were used to determine the extent to which each variable contributed to the prediction of rank in class. PERSUB uses the statistical techniques of multiple linear regression to determine F ratios and exact probability values to four place accuracy. It presents an efficient method for programming computation of the estimated probability for any specific F value. The estimate is based on the actual distribution of scores and may be used with degrees of freedom ranging from four to one thousand (4). The F ratio is computed between a full model regression equation containing all predictor variables under consideration and a restricted model.

In the restricted model in this analysis the information about a particular variable is not included and the predictive efficiency of this equation is compared with the predictive efficiency of the full model's equation, in which all the information for each variable is included. In the equation used in Table 2 the unit vector, SCAT V, SCAT Q, and all Teacher's Ratings are included in an equation as linear combination of vectors to predict to the criterion of R. The predictive efficiency of this equation which includes all variables and is called the full model is given by the R^2 value 0.6794. A restricted model is then constructed holding constant, for example, all other variables except SCAT V. The hypothesis tested is that all members of the sample have the same scores on all variables but different scores on SCAT V and have the same expected rank in class.

By comparing the R^2 of the full model to the R^2 of the restricted model the independent contribution of the variable SCAT V to R can be obtained. Restricted models for the other variables were constructed and tested by comparing their R^2 values to the full model's R^2.

TABLE 2

THE INDIVIDUAL CONTRIBUTION OF TEACHERS' RATINGS AND SCAT SCORES TO PREDICTION OF RANK IN HIGH SCHOOL CLASS

Source	R^2	df	F
Full	0.6794		
SCAT V	0.6689	1/551	18.0474*
SCAT Q	0.6723	1/551	12.2035*
ACCU	0.6631	1/551	28.0165*
COOP	0.6774	1/551	3.4376
E-I	0.6767	1/551	4.6407**
I-L	0.6786	1/551	1.3750
R-R	0.6762	1/551	5.5002*
P-P	0.6780	1/551	2.4063
Self-C	0.6789	1/551	0.8594

Note.--Full model regression equation: SCAT V + SCAT Q + ACCU + COOP + E-I + I-L + R-R + P - P + SELF-C + unit vector predicting to Rank.

*$P < .01$ level of significance.
**$P < .05$ level of significance.

TABLE 3

THE INDIVIDUAL CONTRIBUTION OF TEACHERS' RATINGS OF ACCU, E-I, AND R-R IN PREDICTION TO SCAT V AND SCAT Q

Source	R^2	df	F
Full V	0.2852		
ACCU	0.2749	1/551	7.9402*
E-I	0.2845	1/551	0.5396
R-R	0.2621	1/551	17.8075*
Full Q	0.3205		
ACCU	0.2965	1/551	19.4615*
E-I	0.3163	1/551	3.4057
R-R	0.2978	1/551	18.4073*

Note.--Full model regression equations: ACCU + E-I + R-R + unit vector predicting to SCAT V and SCAT Q respectively.

* $P < .01$ level of significance.

The first multiple linear regression equation included SCAT V and Q and all the teachers' ratings in predicting to the criterion of rank in class. The result of this analysis seen in Table 2 showed that SCAT V and Q and teachers' ratings of ACCU, E-I, and R-R were significant predictors to R ($P < .05$).

A second and third analysis (see Table 3) were then performed in which teachers' ratings of ACCU, E-I, and R-R were used to predict to SCAT V and then SCAT Q. In this analysis teachers' ratings of ACCU and R-R made significant independent contribution to prediction of SCAT V ($P < .01$), while E-I did not. In the third analysis a similar result was obtained in which teachers' ratings of ACCU and R-R predicted to SCAT Q ($P < .01$), and again E-I did not contribute significantly in prediction to SCAT Q.

Male-female differences were also considered a relevant variable in a high-school setting. The PERSUB system was used to compare males and females on each variable separately thus pinpointing male-female differences on those variables which made significant independent contribution in prediction of R. Differences on each variable of importance in scholastic performance could thus be compared for males and females. Males were compared with females on teachers' ratings of ACCU, E-I, and R-R, SCAT V and Q and R. Highly significant differences were obtained for ACCU, E-I, R-R, and R ($P < .01$). For SCAT Q the difference was significant at less than the .05 level of significance but did not reach the .01 level of significance. The comparison between males and females did not reach significance at the .05 level on the other teachers' rating scales. Table 5 presents the means for males and females on these measures. This shows that females are significantly higher in teachers' ratings of ACCU, E-I, R-R, and R (a smaller score indicates a higher rank in class), except on SCAT Q, where males have significantly higher scores than females.

Table 6, which presents the listing of correlations for all variables, indicates that a higher score on any variable, whether SCAT Q, V or teacher ratings (variables 1-9), predicts to a higher rank on the criterion R (variable 10). Since rank in class is the only variable where scaling is reversed, the correlations are negative for R.

DISCUSSION

The results of the analysis shown in Table 2 reveal that SCAT V and Q make significant independent contribution to prediction of R, but teachers' ratings of ACCU, E-I, and R-R are also important in prediction of R. If we rank order the importance of the significant variables in our prediction of R from most important to least important, the following order would result: ACCU, SCAT V, SCAT Q, R-R, and E-I. Thus, the teachers' ratings of accuracy are the most important datum in the analysis with SCAT V and SCAT Q as important predictors also.

The scalar points of ACCU are defined in terms of work performance, going from the low point on scale from 0-25 - "Work is poor. Makes frequent errors," to a C level performance, 55-75 - "Work is well done and reasonably accurate," to the B and A ranges, 80 - 90, with the statement, "Work is of the highest quality." The high degree of prediction from teachers' ratings on this scale probably reflects the teachers' perceptions of how well students are doing in class. Thus, high quality work receives the accolade of an A or a B, and work of lesser quality, less accurate work, receives lower grades. This particular high-school milieu, and probably most academic settings, stress accuracy of response as a criterion of greatest importance in scholastic behavior. Ability factors as would normally be expected are also important. Verbal skills and, next, quantitative skills as measured by the SCAT, have predictive significance in relationship to course content. E-I, which is a measure of industry, again a work-related measure, predicts to R. R-R which may reflect a student's ability to operate conscientiously without immediate supervision, predicts to R and shows how the characteristic of dependability has a positive relationship with grades. The teachers' judgment of a combination of work-related skills and measures of academic ability account for the predictive efficiency of these variables.

TABLE 4

DIFFERENCES BETWEEN MALES AND FEMALES ON ACCU, E-I, R-R, SCAT Q, SCAT V, AND R

Source	R^2	df	F
ACCU	0.1379	1/559	89.3996*
E-I	0.1359	1/559	87.8853*
R-R	0.1318	1/559	84.8830*
SCAT Q	0.0092	1/559	5.1990**
SCAT V	0.0062	1/559	3.4905
R	0.0861	1/559	52.6644*

Note.--Full model regression equation: General form based on binary sectors denoting sex group membership: Male = Female predicting to criteria of interest.

* $P < .01$ level of significance.
** $P < .05$ level of significance.

TABLE 5

MEAN VALUES OF MALES AND FEMALES ON SCAT V AND Q, ACCU, E-I, R-R, AND R
MALES: N = 290 FEMALES: N = 271

	Males	Females
SCAT V	441.127	466.254
SCAT Q	631.534	611.066
ACCU	8.179	8.923
E-I	7.709	9.198
R-R	7.850	9.488
R	350.655	261.693

These results suggest that basic academic skills as measured by SCAT are important, but also work quality, dependability and output are important variables in achievement in a high-school setting. Higher marks are given to work which is done by diligent and conscientious students whose work is accurate, complete, and of high quality.

Students who are helpful and agreeable (COOP), who are independent and self-directing (I-L), and who are self-confident (S-F) are not necessarily those students who obtain high grades. These scales are measuring types of student behavior which are not work-related but reflect positive qualities in personality variables. A high score on dimensions relating to personality dynamics, which is ordinarily regarded in a positive light in a psychological sense, may not necessarily have a significant relationship to success in high-school (see Table 1: Teachers' Rating Scale).

Low scores on these variables would reflect disagreeableness (COOP), timidity (S-F), and the inability of the student to direct himself into any independent line of endeavor (I-L). These types of behavior are perhaps a reflection of cultural differences in that these students are usually reticent since many of them are of Japanese ancestry. This finding then does not disagree with Ausubel and others (1) that teachers' ratings of adjustment were significantly related to grade point averages. A particular type of school environment may be shown by this finding. The results indicate that the primary emphasis within the classroom is placed upon work-related skills and little or no weight is given to personality variables as far as grades are concerned.

Predicting to SCAT V from teachers' ratings of ACCU, E-I, and R-R, we note in Table 3 that R-R has greater predictive significance than ACCU, and E-I does not make a significant independent contribution to prediction of SCAT V. Similarly, with SCAT Q, ACCU and R-R make approximately equal independent contributions to predictive effectiveness, while E-I does not quite reach predictive significance in this instance.

E-I predicts to R, but not to SCAT V and only to a moderate degree to SCAT Q. Though E-I is an important variable in high-school performance, E-I as a measure of industry is not significantly related to verbal aptitude scores. Teachers' ratings of accuracy and reliability may be affected by verbal ability, but this effect does not extend to E-I. E-I may be a separate dimension related to success in high-school, which is independent of general verbal aptitude. It may therefore be a unique and important variable in prediction to academic success in a high-school setting.

On close examination (see Table 3) it may be seen that E-I predicts somewhat better to SCAT Q than to SCAT V. It may be that these three teachers' ratings are also measuring personality characteristics which facilitate performance in quantitative skills. The Japanese in this population, who are characterized by a high need for achievement and, in the female, a significantly higher degree of clinical anxiety, perhaps indicating a high drive state, are superior to the other largest group in this population, the Cosmopolitan or mixed ancestry group in all teachers' ratings, R and SCAT scores. The conscientious, accurate student is probably, then, of Japanese ancestry (3).

Male-female differences indicate that females have significantly higher scores than males on the teachers' rating scale variables which predict to the criterion of R. The list of means shown in Table 4 indicates that only in SCAT Q are males significantly higher than females. The mean differences on SCAT V do not reach the .05 level of significance. Females

TABLE 6

INTERCORRELATIONS BETWEEN SCAT V AND Q, TEACHERS' RATINGS AND R

	1	2	3	4	5	6	7	8	9	10
1 = SCAT V	1.0000	0.6762	0.5197	0.4930	0.4948	0.5383	0.5095	0.4538	0.4455	-0.5758
2 = SCAT Q		1.0000	0.5400	0.5057	0.5132	0.5329	0.5457	0.4638	0.4164	-0.5842
3 = ACCU			1.0000	0.8501	0.8917	0.8787	0.8638	0.7698	0.7412	-0.7736
4 = COOP				1.0000	0.9322	0.8904	0.9246	0.8099	0.7769	-0.7202
5 = E-I					1.0000	0.9125	0.9365	0.8225	0.7535	-0.7630
6 = I-L						1.0000	0.9039	0.7633	0.8101	-0.7517
7 = R-R							1.0000	0.8401	0.7229	-0.7635
8 = P-P								1.0000	0.6399	-0.6816
9 = S-F									1.0000	-0.5987
10 = R										1.0000

Note.--Negative correlations between R (10) and other scores are the result of smaller R values showing higher rank in class.

are significantly higher on R, indicating that even though males have higher quantitative abilities on the average, the females' superiority in teachers' ratings of ACCU, E-I, and R-R are the critical variables which differentiate between male versus female performance in a high-school setting. This comparison of male-female differences was performed on each variable in a separate analysis for those variables which made an independent contribution to prediction of R.

The significantly better performance (R) of females in a high-school setting, even with a disadvantage in quantitative abilities, as shown in Table 3, probably relates to a personality variable. This may be accounted for by the fact that ACCU, E-I, and R-R are work-related values involving accurate, precise, conscientious work. The difference between male and female performance suggests that females behave in a manner which conforms to the values judged by teachers as being important to scholastic endeavors. These results suggest that the female role in a high-school setting of this type is a more conforming one than that of her male classmate. Sex grouping, as well as SCAT V and Q, and teachers' ratings relating to the quality and regularity of work output, is therefore an important variable as predictor of rank in class in high school.

Academic performance in high school is not the sole product of scholastic aptitude but is best explained by a combination of scholastic aptitude and a constellation of personality traits which affect students' performance in a classroom situation. Prediction based solely on standardized tests may omit salient variables in prediction to academic success in high school. These findings suggest that school-generated rating schedules may have predictive validity, especially where these scales are teachers' ratings of work-related skills. Improved prediction of academic success in a high-school setting may, it is suggested, be obtained through the use of well constructed bipolar teacher rating scales in conjunction with standardized measures of scholastic aptitude such as the SCAT.

SUMMARY

The records of 561 high-school seniors were examined for SCAT scores, teachers' ratings, and rank in class. Highly significant F ratios were obtained for SCAT V and Q, teachers' ratings of ACCU, E-I, and R-R in their prediction to rank in class. R-R and ACCU were found to be significant predictors of SCAT V and SCAT Q, but E-I did not predict significantly to either. Females were higher on teachers' ratings of ACCU, E-I, and R-R but were not significantly different from males on SCAT V and were significantly lower than males on SCAT Q. This indicates that the student who does well in high school is the student who does precise and accurate work, who is industrious, and who can perform well without supervision. These findings also suggest that the industrious student does better in high-school setting, but that industriousness is not associated with SCAT V and is only associated to a moderate degree with SCAT Q. E-I as a measure of industriousness is then a unique and important variable in prediction to academic success in high school.

Females were higher on teachers' ratings which predicted to rank in class. They were not superior to males on SCAT V and were significantly below males on SCAT Q. The results suggest that much of the variance in high-school rank may be accounted for by personality factors relating to work requirements in a high-school setting. Since females were higher on ACCU, E-I, and R-R, personality factors that relate to conformity to teacher expectations may account for their superiority in high-school performance as measured by rank in class.

FOOTNOTE

1. This research was supported by NSF funds administered by the Research Council of the University of Hawaii.

REFERENCES

1. Ausubel, D. P., Schiff, H. M., and Zeleny, Marjorie P., "Validity of Teachers' Ratings of Adolescents' Adjustment and Aspiration," Journal of Educational Psychology, 45:394-406, 1954.

2. Bottenberg, P. A., and Ward, J. H., Applied Multiple Linear Regression, Technical Documentary Report PRL-TDR-63-6. Defense Documentation Center, Defense Supply Agency, Washington, March 1963.

3. Dixon, P. W., Fukuda, Nobuko K., and Berens, Anne E., "The Influence of Ethnic Grouping on SCAT, Teachers' Ratings and Rank in High School Class," Journal of Social Psychology, 75:285-286, 1968.

4. Neel, G. I., Estimation of Probabilities Associated With the F Statistic by Digital Computer Techniques, Technical Documentary Report PRL-TDR-63-7. Defense Documentation Center, Defense Supply Agency, Washington, March 1963.

5. Osgood, C. E., Suci, G. J., and Tannenbaum, P., The Measurement of Meaning, University of Illinois Press, Urbana, 1957.

6. Rugg, H. O., "Is the Rating of Human Character Practicable?" Journal of Educational Psychology, 13:30-40, 81-93, 1922.

7. Symonds, P. M., Diagnosing Personality and Conduct, Appleton-Century Publishing Company, New York, 1931, 602 pp.

8. Williamson, R. G., and Cole, C., "Factors in Scholastic Performance: the Behavior Differential," Personnel and Guidance Journal, 44:962-965, 1966.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

THE EFFECTS OF COMPUTER BASED RESOURCE UNITS UPON INSTRUCTIONAL BEHAVIOR

GEORGE S. HOLDEN

Northern Illinois University

ABSTRACT

Computer based resource units were developed to improve the individualization of instruction through unit teaching. Instructional behaviors of teachers were examined to see if this kind of computer aid brought about a change in the way teachers teach. A pre-post two group experimental design was used in which the data were collected by a group of trained observers using an instruction observation tool. Pre-post mean scores were tested and found to be significant in five out of seven dimensions of instructional behavior. Variety and kinds of instructional materials and methods were also examined. The data indicate that teachers who use computer based resource units increase the number and improve the quality of many individualized instruction tasks.

THE REVOLUTION which professional education has been undergoing the last few years involves changes in just about every facet of education. Included are reexamination of educational objectives, restructuring of subject matter, new practices in staff utilization and organization, originality in building design and function, diversification of instructional materials and methodology, and refinement of measuring devices. Most of the changes that are taking place seem to be related in some way to an increased awareness of the need to be in touch with the person of the individual in the classroom.

The alarming rate of change, along with the move toward the individualization of instruction, has served to complicate and confuse the decision making which must be done by elementary and secondary school personnel. Computer based resource units, which have been given the name "Resource Guides," were developed to help the teacher make decisions in his complicated planning tasks.[1], [2] Basically, in this study the computer serves as a retrieval system designed to assist the teacher in his decision making about classroom objectives, subject matter content, instructional activities, materials, and measuring devices. The resource guide was designed primarily to help the teacher individualize instruction through unit teaching, and it was necessary to see if the instructional tasks in which teachers engage become more individualized as a result of using the computer materials. Therefore, this article reports on a study which attempted to get at the question: Does the use of resource guides cause a varied selection of teachers to more nearly individualize their instruction by (1) encouraging pupils to engage in independent thinking, (2) creating an accepting atmosphere in the classroom, (3) making appropriate selection and use of instructional materials, (4) making appropriate selection and use of teaching methods, (5) motivating pupils through challenge without threat, and (6) sensing the needs of individuals in the classroom?

To get at the question, a pre-post measure of instructional behavior was recorded so that observable differences could be noted and objectified. To discover whether or not these differences were statistically significant, the following null hypothesis was tested:

> There is no significant observable difference between the regular instructional behavior of a varied selection of teachers and the instructional behavior of the same teachers recorded at a time when they are following plans that were developed from computer based resource guides.

DESIGN AND METHODS

In an attempt to investigate the effects of computer-based resource guides upon the instructional behavior of a varied sample of classroom teachers, nineteen high school teachers and nineteen third grade teachers were each observed four times by trained observers using Jason's Instruction Observation Record (IOR).[3] The first two observations, representing the pre-data of a pre-post design, were implemented under essentially the same conditions for all of the teachers.

For the post-observations, the teachers were alternately assigned to experimental and control groups. The control group was used to guard against teachers changing instructional behavior because of:

(1) the Hawthorne effect: experimental teachers were asked to use the computer based resource guide, while the control teachers were

asked to use materials drawn from their usual sources.

(2) an increased awareness of their own objectives; both experimental and control teachers were asked to identify the objectives they wished to achieve.

(3) an increased awareness of individual pupil characteristics; both experimental and control teachers were asked to complete a pupil characteristics form for each pupil in his class.

A further function of the control group was to guard against biased reporting of change in instructional behavior. Observers were not told whether they were observing experimental or control group teachers.

Even though the training of the observers produced a high level of consistency among the observers in the results they achieved when they used the observation instrument, each teacher in the study had only one observer assigned to him. That is, each observer was responsible for making all four observations of each teacher assigned to him. The instructional behaviors which were observed and rated were those behaviors represented by the seven scales of the IOR.

Scale A-Attitude to Difference
Scale B-Sensitivity to Physical Setting
Scale C-Attitude to Students
Scale D-Use of Instructional Materials
Scale E-Reaction to Student's Needs
Scale F-Use of Teaching Methods
Scale G-Use of Challenge

At the end of the second visit, the observers left a listing of forty to sixty objectives that might be applied to the particular unit in question and instructed each teacher to select not more than ten that he would like to have his class achieve as a whole, and any four of these ten for each pupil in the class. The observers also requested each teacher to complete one pupil characteristic check list for each pupil in the class. This check list covered the following characteristics:

1. I. Q.
2. Social Class
3. Reading Level
4. Sex
5. Grade Level
6. Interests
7. Vital Statistics (family information)
8. Handicap-Behavior
9. Developmental Tasks

The objectives and pupil characteristics were fed into the computer, which sorted through vast amounts of coded material about the unit being studied. The result was the production of a resource guide - a reservoir of materials containing the following data relevant only to the particular class for which the resource guide was generated.

Part A - 1. Subject matter outline
2. Large group activities
   a. Introductory
   b. Developmental
   c. Culminating
3. Small group activities
4. Instructional materials
5. Measuring devices

Part B - 1. Individual activities per each pupil per each of four objectives
2. Instructional materials per each pupil per each of four objectives.

Significance of difference between the pre and post samples and between the experimental and control mean differences were tested at the .05 level by using a t-test.

THE FINDINGS

The carefully balanced scales of the IOR provide for the heuristically useful practice of quantifying each instructional behavior in digital symbols. The symbols range from one, representing extreme avoidance of individualized instruction tasks, to twenty, representing major involvement in activities designed to further the individualization of instruction.

Table 1 reveals that all of the observed differences for the experimental group were in the hypothsized direction of greater individualization of instruction and that six of the eight changes in instructional behaviors were statistically significant.

TABLE 1

MEAN SCORES FOR THE TOTAL EXPERIMENTAL GROUP

Scale	A	B	C	D	E	F	G	X
Pre	12.9	13.7	13.7	14.1	12.5	13.1	12.1	13.3
Post	14.7	15.6	15.7	16.5	13.9	15.2	14.7	15.3
Change significant at .05 level	+1.8 *	+1.9	+2.0 *	+2.4 *	+1.4	+2.1 *	+2.6 *	+2.0 *

You will note that although scale B showed a greater mean gain score than scale A, the former was not statistically significant while the latter was. Each scale of the IOR provides for the observer to respond to "Insufficient Evidence" or "Inappropriate for this Session" as well as to the scaled behavior section. On scale B, twenty-four of the thirty-eight teachers in a sample were not given a scaled rating for either the before behavior, the after behavior, or both. So the failure to reach significance on scale B is in part a function of the number in the sample.

The failure of scale E to reach significance is another matter. It might possibly be due to the nature of the behavior rather than to an innocuous quality of the Resource Guide. Other instructional behaviors have a control-by-will quality in them. That is, it is more or less a simple matter of professional decision-making to engage in the other instructional behaviors. The teacher has a more or less direct control on these; but not so with sensitivity to students' needs. Students' needs are many times represented by subtle behaviors. The teacher quality looked for here is a skill in using a silent language—a skill in FEELING communication from others. A pupil who squints his eyes, another who stares out the window, or one who talks to a classmate during a lecture—each one is telling the teacher something. More than this, a pupil who asks a question of the teacher may not be verbalizing his real question at all. His real question might go something like this: "Am I doing all right, teacher?" or "Teacher, do you like me?" The effective teacher must be capable of exercising empathy—of being sensitive to a silent language—a language that is conveyed on the vehicle of feeling and employs symbolizations of emotions, gestures, bodily movements and facial expressions. Even time and space are important indicators of communication. Silence can indicate thought, fear, examination of a new idea, hostility, assent, agreement or disagreement and many more qualities of communication. Teachers must be capable of sensing this "feeling" communication.

This is not the kind of ability one can call upon if he has never been sensitized to it. It is a skill that one might develop over a period of time, but does not have the control-by-will quality identified in the other instructional behaviors.

The pre-post change for the control group failed to reach significance on any of the scales except Scale B. Here again, the sample for this scale was inadequate.

It would be of no consequence if the gain scores of the experimental group registered significance and the gain scores of the control group failed to reach significance unless the difference between the gain scores were found to be significant. Non-significant difference between gain scores could be found if both groups registered plus charges but the control group just missed reaching significance and the experimental group just made it. Under these conditions the post-scores could be so close together that even though one was significant and the other was not, the difference between the two groups would not be a real difference. To be sure that this did not occur, the differences between the control and experimental gain scores were computed and found to be significant at the .05 level on all scales but Scale B.

The difference between the effect of the resource guide upon elementary and high school teachers was also investigated and, indications are that high school teachers are affected more. The data indicates that the reason for this seems to be that elementary teachers begin with more of an individualized instruction orientation and behavior.

A quantitative analysis of methods and materials used and of the number of pupil questions, comments and interactions supports the data reported. The experimental group post sessions used more materials, a greater variety of teaching methods and elicited more pupil interaction.

CONCLUSIONS

The following conclusions have been drawn within the limitations of this study:

1. There is a difference between the regular instructional behavior of a varied selection of teachers and the instructional behavior of the same teachers recorded at a time when they are following plans that were developed from computer based resource guides.

2. Resource guides significantly affect some dimensions of instructional behavior but fail to produce significant change in other dimensions of teaching. The significant changes in instructional behavior of teachers who use resource guides are changes that increase the number and improve the quality of the following individualized instruction tasks:

   A. encouraging pupils to engage in independent thinking

   B. creating an accepting atmosphere in the classroom

   C. making appropriate selection and use of instructional materials

   D. making appropriate selection and use of teaching methods

   E. motivating pupils through challenge without threat

   F. employing a wider variety of instructional materials

   G. using a greater number of individual and small group methods of teaching and fewer large groups methods

   H. encouraging more pupil involvement and interaction

3. Resource guides fail to produce significant changes in teachers' sensitivities to the needs of pupils or in their sensitivities to the effects of the physical setting.

SUMMARY

The results of the present study indicate that the development of the resource guide may contribute significantly to curriculum planning and to the improvement of individualization of instruction through unit teaching. The resource guide has been effective in causing a varied selection of elementary school and high school teachers to make observable changes in their instructional behavior, and these changes are in the direction of individualized instruction.

A persistent concern of educators has been to find ways to combat the tendency of a mass education in our conforming and success oriented school society to dehumanize the individual. It seems paradoxical that an answer might be found in a computer, which is another object of accusation in the fight against losing sight of the individual. Yet, the computer-based resource guide seems to provide assistance in maximizing regard for the individual in the classroom.

REFERENCES

1. Eilsele, James E., "Computers In Curriculum Planning," Educational Technology, November 30, 1967.

2. Harnack, Robert S., "Computer Based Resource Units," Educational Leadership, December 1965.

3. Jason, Hilliard, An Analysis of Teaching Practices at Seven Selected American Medical Schools, unpublished doctoral dissertation, State University of New York at Buffalo, 1961.

New Title From DERS

RESEARCH AND DEVELOPMENT TOWARD THE IMPROVEMENT OF EDUCATION

Edited by Herbert J. Klausmeier and George T. O'Hearn. 1968

176 pages **$5.75 softcover**

In this book 20 eminent scholars, researchers, and research administrators present their latest thinking about the processes and conditions of human learning, the processes and programs of instruction, and the application of research and devlopment strategies to educational problems.

As the editors point out in the Preface, the application of research and development strategies for the improvement of educational practices is a pioneering venture begun in 1964 under the provisions of the Cooperative Research Program of the United States Office of Education and now continuing through Title IV of the Elementary and Secondary Education Act of 1965. Working from a proposed model outlined in Chapter 16, the editors have organized the book around the major components of an instructional system.

Basic to the improvement of education is knowledge about the processes and conditions necessary for efficient learning. In Part I, four authors discuss retention and recall, thought processes, creative thinking, and motivation. A major component of an instructional system is subject matter content and sequence. In Part II, five authors provide scholarly insights into the structures of such disciplines as mathematics, science, English, and reading. They state or imply that the structure of the discipline itself, as formulated by one or more scholars, has provided much of the basis for recent curriculum devlopment.

Instructional materials and media can be considered as the interface between the learner and the subject matter being learned. In Part III, the discussion focuses on instructional television, on some aspects of the status, defects and needs of instructional research, and on computer based instruction. Teaching methods, or the interactions which occur between a teacher and students, have been the subject of many research studies in the last decades. In Part IV, two recent paradigms for research on teaching are presented. One approach discusses what is commonly called micro-teaching and the other is interaction analysis.

Finally, in Part V, three research administrators discuss the current application of research and development strategies for the improvement of educational practices. The present scene in educational R & D is one of rapid change brought about largely by the advent of federal funding. The development of these changes with reference to federal involvement is traced in one chapter. In another chapter the focus and programs of the Wisconsin Research and Development Center for Cognitive Learning are described. And in the final chapter, an author presents an output oriented model of R & D.

Dembar Educational Research Services, Inc. Box 1148
Madison Wisconsin 53701

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

PREDICTING GRADE ONE READING PERFORMANCE: INTELLIGENCE VS. READING READINESS TESTS[1]

KENNETH D. HOPKINS
University of Colorado

E. GEORGE SITKEI
Los Angeles County Schools

ABSTRACT

Since both intelligence and reading readiness tests are widely used to predict reading success in grade one, both types of tests were administered to all entering first-grade pupils in two schools. Performance on these measures was correlated with end-of-year teacher marks and scores on a standardized reading test. Although the predictive validity coefficients for the readiness test were larger than those for the intelligence test on both criteria, the differences were not statistically reliable. When IQ was added to reading readiness scores, the multiple correlation rose from .61 to .67. Adding father occupation, sex, and age failed to meaningfully increase the accuracy in prediction; the multiple R's increased only to .68.

THE COMPARATIVE validities for predicting grade-one reading success of tests of intelligence and reading readiness is an issue that has received surprisingly little attention, especially in view of the fact that the functions of both types of tests are highly similar for beginning first-grade pupils. Several studies (e.g., Dobson and Hopkins (1)) have been reported in which one type of test was used, although noncomparable samples make an evaluation of the comparative efficacy uncertain. Mattick (4) administered both types of tests but used a criterion of success obtained after only two months of grade one. In addition, different schools employed different tests, creating a special problem, since some schools were ungraded. This factor would have large effects on obtained correlations since variability within such classes would be considerably reduced. Fry (2) obtained inconsistent results in predicting various reading criteria, although the pooling of information from different experimental methods makes a generalization of his results equivocal.

PROCEDURE

All entering grade-one pupils in two elementary schools in a lower- middle class community were administered both the Lee-Clark Reading Readiness Test (LCRR) (1962 revision) and the California Test of Mental Maturity (CTMM) (1957 S-form) during the first three weeks of school, the latter test always being administered last. Performances on these tests were correlated with two independent criteria, subsequently gathered near the end of the school year: (a) scores on the Lee-Clark Reading Test: Primer (LCRT) (Form A), and (b) end-of-year teacher marks, using a 4-category scale. In addition to the two predictors and two criteria, other pupil data were collected in order to evaluate the degree of improvement that could result from the use of multiple predictors. The teachers reported their final marks prior to being made aware of the pupil's scores on the standardized reading test. Over-aged pupils were removed from the sample to avoid the confounding effects of non-promotion, etc. (Age in months: $\overline{X} = 76.8$, $s = 4.1$). Since a Type I error in this situation is quite innocuous, α was set leniently at .10 in order to reduce the probability of a Type II error, the more serious in this type of study.

RESULTS

Table I gives the predictive valiadities of the LCRR, language, non-language, and total CTMM IQ's for the 157 pupils on whom complete data were available.

Although the predictive validity coefficients for the reading readiness test numerically exceeded those for the intelligence tests on both criteria, the differences proved to be insufficiently great to allow rejection of the null hypothesis for either criterion. In predicting teacher marks, correlation coefficients of .571 and .513 and corresponding .95 confidence intervals of .455 - .668 and .386 - .619 were observed for the reading readiness and intelligence tests respectively. The difference between the non-independent correlation coefficients was not significant ($t = .90$). For predicting performance on the standardized reading test, corresponding non-significantly different coefficients of .612 and .541 ($t = 1.16$) were observed, the larger again being associated with the readiness test. The .95 confidence intervals were .503 - .701 and .446 - .642 respectively.

A multiple regression using both IQ and LCRR

TABLE 1

VALIDITIES OF VARIOUS TESTS FOR PREDICTING GRADE ONE READING PERFORMANCE

	Predictors					Criteria	
	(1)	(2)	(3)	(4)	(5)[b]	(6)	(7)
Predictors							
(1) Lang. IQ		.476	.853[a]	.364	.342	.448	.497
(2) N-Lang. IQ			.857[a]	.442	.422	.431	.430
(3) Total IQ				.466	.445	.513	.541
(4) LCRR (rs)					.934	.571	.612[c]
(5) LCRR (gp)						.595	.590
Criteria							
(6) T-marks							.748
(7) LCRT (rs)							
$\overline{X}$	101.8	103.4	102.8	49.0	1.00	2.45	26.0
s	16.8	17.0	14.7	10.3	.55	.94	8.1

[a] Spurious correlation since total IQ is an average of the language and non-language IQ.

[b] Both raw scores and grade placement units are reported to illustrate the essential interchangeability of findings irrespective of which of these two scales is used.

[c] Since age is, in effect, held constant in IQ determinations, $r_{47.\text{age}}$ was computed to make this value also independent of age effects. The partial correlation did not differ from the value given in the table until the fourth decimal place.

as predictors yielded multiple correlations .05-.06 (corrected for bias) greater than the LCRR alone. Adding father occupation, sex, and age to the regression failed to add meaningfully to the accuracy in prediction; the multiple R's only increased approximately an additional .01 in value.

In view of the fact that the reading readiness test did at least as well in predicting first-grade reading performance, it was considered to be preferable to the intelligence test when other relevant factors are considered: (a) it requires considerably less testing time, (b) it is more easily and meaningfully interpreted, (c) the effects of improper interpretation are much less serious to the pupil, and (d) it is less expensive.

The primary conclusion of the study, that the readiness test did at least as well in predicting first-grade reading performance, is consistent with those reported by Mattick (4) for the Metropolitan Reading Test and the Lorge-Thorndike Intelligence Test; it is also in agreement with Hahn's (3) study in which the Murphy-Durrell Reading Readiness Analysis and the Metropolitan Readiness Test were compared with the Pitner-Cunningham General Ability Test: Verbal Series.

FOOTNOTES

1. The cooperation of Miss Mary F. Sanders and Mr. Robert E. Oaks, principals in the Pomona Unified School District, in conducting this study was greatly appreciated.

REFERENCES

1. Dobson, J. C., and Hopkins, K. D., "The Reliability and Predictive Validity of the Lee-Clark Reading Readiness Test," Journal of Developmental Reading, 6:278-81, 1963.
2. Fry, E., "Comparisons of Three Methods of Reading Instruction," U. S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research. Project No. 3050, 1966.
3. Hahn, H. T., "Three Approaches to Beginning Reading Instruction--ITA, Language Arts and Basic Readers," The Reading Teacher, 19:590-94, 1966.
4. Mattick, W. E., "Predicting Success in the First Grade," Elementary School Journal, 63: 273-76, 1963.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

SEMANTIC DIFFERENTIAL FACTOR STRUCTURE WITH CONCEPTS AND SUBJECTS FROM EDUCATION [1]

ROBERT B. KANE

Purdue University

ABSTRACT

This research sought answers to three questions: Are factors discernible as evaluation, potency, and activity evident? Do scales which have been associated with one of these factors continue to register high and relatively pure loadings on that factor? Do adjectives traditionally labeled "positive" maintain this posture with educational concepts? Semantic Differential Scales were administered twice to elementary education majors. The resulting matrixes were factor analyzed and rotated orthogonally. Factor I was evaluation but included two scales often associated with other factors. Factor II seemed to be intellectual rigor, rather than potency. Factor III seemed to be an activity dimension. In about one-half of the cases "feminine" defines the positive end of the masculine-feminine scale. Extrapolating an SD scale's factorial content or its polarity from various contexts to elementary education seems unwarranted. Factor analysis should be the first step in each SD data analysis.

SEMANTIC DIFFERENTIAL (SD) bipolar adjective scales exhibit stability across a wide variety of S's, but they do not exhibit comparable stability across concepts — Shaw (10), Osgood, Suci and Tannenbaum (8), Husek and Wittrack (3), Osgood (9), Hartman (2), Tanka and Osgood (11), Ohnmacht (7). Thus, while subject-scale interaction is low, concept-scale interaction tends to be high. None the less, the three classic SD factors labeled evaluation, potency, and activity generally appear and usually account for 50 percent or more of the total variance.

Osgood, Suci, and Tannenbaum (8), cautioned that the meanings of scale-defining adjectives may change from concept to concept and that the assumption of fixed polarity for a scale may not be tenable as concepts are varied. Moreover they reported that the activity factor fared poorly in single concept factor analyses. Among nineteen concepts tested, activity was indentifiable as a factor in only eight; it was distributed among other factors for various concepts.

The problem was to determine the factorial structure of SD scales that have become accepted widely as paradigms of the three factors termed evaluation, potency, and activity when they are used by prospective elementary teachers to rate educational concepts. In particular answers to the following questions were sought:

1. Are factors discernable as evaluation, potency, and activity evident?
2. Do scales which have become associated with one of the three factors listed above continue to register high and relatively pure loadings on their respective factors?
3. For each scale does the adjective traditionally associated with the positive end of the scale maintain this posture when the scale is used to rate educational concepts?

METHOD

Seventeen concepts were selected that are directly related to classroom activities in the elementary schools. Nine of them related to the major curricular areas: they were language arts, mathematics, science, social studies, teaching children, teaching children language arts, teaching children mathematics, teaching children science, and teaching children social studies. The remaining eight concepts consisted of the defining sentences of Flanders' (1) categories of teacher behavior: The teacher criticizes or deprecates pupil behavior with intent to change it. The teacher gives directions or orders. The teacher expresses or lectures about her own ideas. The teacher asks questions to orient pupils to school work. The teacher asks questions to stimulate pupil participation in decision making. The teacher accepts, clarifies, and supports the ideas and feelings of pupils. The teacher praises or encourages pupils. The teacher justifies his own position or authority. Each concept was rated on 14 scales which were selected by reviewing the literature for SD scales which consistently exhibited high and relatively pure factor loadings across a variety of concepts judged by different kinds of subjects. Evaluative scales selected were good-bad, nice-awful, positive-negative, heavenly-hellish, optimistic-pessimistic, and happy-sad. Potency scales selected were strong-weak, heavy-light, hard-soft, and masculine-feminine. Activity scales included fast-slow, active-passive, hot-cold, and difficult-easy.

The S's were seventy-one seniors in elementary education. The SD's were administered in an educational foundations course. Each S responded to the questionnaire both before and after an 8-week student teaching period. (This study is a part of research which was designed to measure attitude changes that develop during student teaching. Reporting factor structures for both pre- and post- student teaching administrations in this report is done only to increase the data available for analysis). Four random orders of concept presentation were used. The scale presentation order was selected randomly, and the order of "positive" and "negative" adjectives within scales was randomized. Each questionnaire included directions to S as suggested by Osgood, Suci, and Tannenbaum (8). The S's were given ample in-class time to complete the questionnaire, and every S completed every item.

Thirty-four 14 x 14 matrices of product-moment correlations were produced. Each of these was factored using principal components analysis. Unities were used to estimate communality, and each analysis was followed by an orthogonal rotation to Kaiser's (4, 5) Varimax criterion. Linear correlations were justified because no systematic nonmonotonicity was observed among variables in the several matrices. While nonlinear relations undoubtedly exist among the variables, a linear correlation model yields a reasonable measure of the degree of relationship for a monotonic relation.

RESULTS

The Varimax criterion terminated the rotation after three factors in twelve of the thirty-four analyses, and in no case was the rotation of more than five factors necessary. The proportion of total variance accounted for by the first three factors ranged from 0.462 to 0.663. When fourth and fifth factors were rotated, they appeared to be reoccurrences of heavy loadings on evaluative scales, or a factor which more frequently appeared as Factor II or III but was deposed in that particular analysis, or they seemed uninterpretable. This report will, therefore, be restricted to an analysis of the factor loadings for the first three rotated factors.

Factor I is quite clearly the evaluative demension^2. All six of the scales chosen because they exhibited high and relatively pure loadings on the evaluative dimension in other studies yielded loadings ≥ 0.30 in 29 or more of the 34 factor analyses summarized here. In addition the scale active-passive qualified in 32 of the 34 cases and the scale strong-weak qualified in 31 of the 34 cases.

Factor II is best defined by the scales hard-soft and difficult-easy. In only four cases out of 34 does one of these scales meet the 0.30 criterion while the other scale does not. Moreover, in 19 cases out of the 30 in which they are paired they ranked first and second.

Factor III is defined by the following scales: fast-slow, masculine-feminine, and hot-cold. In 11 cases out of 34 all three scales meet the 0.30 criterion. In an additional 11 instances, two out of the three met the criterion.

TABLE 1

FREQUENCY OF OCCURRENCE OF FACTOR LOADINGS ≥ 0.30 ON FACTORS FOR WHICH SCALE IS NOT LISTED

	Relative frequency with which scale loads ≥ 0.30 on other factors
Factor I scales	
happy-sad	0.18
active-passive	0.16
good-bad	0.07
heavenly-hellish	0.43
positive-negative	0.09
nice-awful	0.16
optimistic-pessimistic	0.16
strong-weak	0.21
Factor II scales	
hard-soft	0.07
difficult-easy	0.24
Factor III scales	
fast-slow	0.37
masculine-feminine	0.12
hot-cold	0.15

Data in Table 1 provide information relative to the factorial "purity" of the high-loading scales for Factors I, II, and III across the set of educational concepts studied.

If we were to reject all scales with loadings ≥ 0.30 on other factors in 25 percent or more of the cases, then *heavenly-hellish*, *fast-slow*, and *hot-cold* would be eliminated.

CONCLUSIONS

Active-passive, the traditional paradigm of the activity factor, is clearly an evaluative scale with these educational concepts and subjects. In only three cases out of 68 does it load above 0.50 on Factor II or III. Similarly *strong-weak*, a traditional potency scale, becomes an evaluative scale. In only two cases does it load above 0.50 on Factor II or III. Factor II is a hybred of the traditional activity and potency factors. In light of the concepts rated and the factor loadings it appears that *hard-soft* and *difficult-easy* become synonomous, and that Factor II is associated with intellectual rigor, profundity, substance, and opaqueness rather than physical hardness, strength, or the like. The shift in connotative meaning among Factor II scales may account for the *masculine-feminine* scale's low loadings on this factor. While *masculine-feminine* may link with *hard-soft* for physical attributes, it certainly does not form such a link for intellectual attributes connected with these educational concepts. In these settings it is linked with two traditional activity scales in Factor III.

While *masculine-feminine* is the "purest" scale in Factor III, it does not maintain a stable polarity across the concepts included. An indication of this may be observed in Table 2.

TABLE 2

COMPARISON OF THE POLARITIES OF PAIRS OF SCALES WHOSE LOADINGS WERE ≥ 0.30 ON FACTOR III

	Same Polarity	Opposite Polarity
masculine-feminine with fast-slow	10	6
masculine-feminine with hot-cold	8	10
fast-slow with hot-cold	27	2

These data suggest that in about one-half of the cases 'feminine' defines the positive end of the scale. Considering the predominance of female teachers in elementary school classrooms, this outcome is not surprising.

It seems clear that some scales which have come to be regarded as activity or potency scales because of their relatively consistant performance in many studies did not perform in expected ways with these educational concepts and subjects. Kerlinger (1964) quotes Osgood as suggesting that SD's should always include scales of known factorial content. This study emphasizes that it is dangerous to suppose that a scale has a "known" factorial content that can be assumed at the outset. Moreover, one scale's polarity switched from concept to concept for the same subjects during a single SD administration. Erroneous results and conclusions would be generated by scoring and analyzing responses based on assumed scale performance.

The 34 factor analyses including rotations required only 3.4 minutes on an IBM 7094 computer. Those who use the SD with educational concepts should perform factor analysis as a first step in data analysis.

FOOTNOTES

1. The work reported herein was performed pursuant to a grant from the U. S. Office of Education, Department of Health, Education, and Welfare.
2. Loadings ≥ 0.30 for factors I, II, and III are presented in the Final Report, Project No. 7-E-053, Grant No. OEG-1-7-070053-1908, *Use of the Semantic Differential Technique to Measure Prospective Elementary School Teacher Attitude Toward Mathematics and Other Subjects*, February, 1968, USOE, Bureau of Research, pp. 11-18.

REFERENCES

1. Flanders, N. A., *Teacher Influence, Pupil Attitudes, and Achievement*. University of Minnesota, Minneapolis, 1960, (USOE Coop Res Project No. 397), (mimeographed).
2. Hartman, F. R., "Behavioristic Approach to Communication," *AV Review*, 11:155-190, 1963.
3. Husek, T. R. and Wittrock, M. C., "The Dimensions of Attitudes Toward Teachers as Measured by the Semantic Differential," *Journal of Educational Psychology*, 53:209-213, 1962.
4. Kaiser, H. F., "The Varimax Criterion for Analytic Rotation in Factor Analysis," *Psychometrika*, 23:187-200, 1958.

5. Kaiser, H. F., "The Application of Electronic Computers to Factor Analysis," Educational and Psychological Measurement, 20:141-151, 1960.

6. Kerlinger, F. N., Foundations of Behavioral Research, Holt, Rinehart and Winston, New York, 1964.

7. Ohnmacht, F. W., "Some Dimensions of Meaning of the Concept Televised Instruction," Educational and Psychological Measurement, No. 2, 26: 395-401, 1966.

8. Osgood, C. E., Suci, G. J., and Tannenbaum, P. H., The Measurement of Meaning, University of Illinois Press, Urbana, 1957.

9. Osgood, C. E., "Studies on Generality of Affective Meaning Systems," American Psychologist, 17:10-28, 1962.

10. Shaw, D. R., Variation in Inter-scale Correlation on the Semantic Differential as a Function of the Concept Judged, unpublished Master's Thesis, University of Illinois, 1955.

11. Tanka, Y. and Osgood, C. E., "A Cross-cultural and Cross-Concept Study of the Generality of Semantic Spaces," Journal of Verbal Learning and Verbal Behavior, 2(5-6), 392-405, 1963.

ON TO THE CLASSROOM

A Guide for Preschool Teachers in Head Start-Type Programs of Compensatory Education

EDITED BY

Robert E. Clasen

200 pages $7.25 Hardcover and $5.75 Softcover

ON TO THE CLASSROOM deals with typical problems common to teachers of disadvantaged preschool children and contains unique suggestions for understanding and meeting the needs of these youngsters. The chapters are based on papers by well-qualified professors and professionals from the preschool education field which were originally presented to a group of Head Start teachers needing help in the various areas covered. The editor says, "Since these works were extremely useful to one group of teachers, they should be useful to others."

The book begins with a chapter which defines "culturally deprived" and offers a frame of reference for the thoughts and ideas presented in the remainder of the book. Each chapter was selected by Dr. Clasen on one criterion: *Does it contain information which our experience has shown that teachers need?* The chapters speak for themselves:

Creating a Learning Environment (numerous hints are given on how this learning climate can be created) (Chapter 2)

The Teacher, The Child and Head Start (the needs of children and a teacher's awareness are discussed) (Chapter 3)

Speech Language Acquisition and Language and Head Start (deal with language diagnosis and teaching strategies) (Chapters 4, 5)

From a Teacher's Point of View (a humorous and heart-rending day to day account of organizing, canvassing, and parent programming in Head Start, plus the happenings in a Head Start classroom from the first class day to the last—all taken from a teacher's log with her commentary and suggestions) (Chapter 6)

A Conversation with a Head Start (A.D.C.) Mother (reveals what the mother of a Head Start child experiences) (Chapter 7)

Programming for Parents (offers surprising views on what this is all about) (Chapter 8)

A Statement by Dr. Clasen summarizes the real purpose of ON TO THE CLASSROOM: "The fondest hope of each of us is that an idea shared through this medium may stimulate a change in a teacher's behavior for the benefit of a child."

PLEASE SEND ME THE INDICATED NUMBER OF COPIES OF: ____________

DEMBAR EDUCATIONAL RESEARCH SERVICES, INC.

POST-OFFICE BOX 1148 • MADISON, WISCONSIN 53701

NAME ____________

ADDRESS ____________

CITY ____________ STATE ________ ZIP ________

☐ I enclose a check for postpaid books ☐ Bill me and I'll pay the postage

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

ACADEMIC ATTITUDES AND PERFORMANCE AS A FUNCTION OF DIFFERENTIAL GRADING SYSTEMS: AN EVALUATION OF PRINCETON'S PASS-FAIL SYSTEM[1]

MARVIN KARLINS[2], MARTIN KAPLAN, and WILLIAM STUART
Princeton University

ABSTRACT

An empirical investigation of the Princeton University pass-fail system is presented. Findings based on questionnaire and academic performance data indicate that the pass-fail grading alternative is not the panacea that some educators envisioned. It is suggested that further study of the pass-fail system be undertaken *before* it becomes a permanent grading procedure in our Universities.

IN THE PAST decade many American educators have displayed increasing hostility toward the competitive (A-F) grading system prevalent in most schools. These critics base their opposition on several assumptions, two major ones being: (a) students will learn more in their courses when the deleterious effects of competitive grading are reduced or eliminated; (b) encouraging learning through incentives such as grades is not in keeping with contemporary educational philosophy which says a student should study because he is interested in the subject matter, not a grade in the subject. These educational assumptions are not without theoretical support and, indeed, are congruent with current psychological notions about the *intrinsic* motivational correlates of learning.

The current unrest of those individuals opposed to the A-F grading system has been of sufficient intensity to bring about a partial de-emphasis on competitive grading in several major universities. Princeton, Rutgers and Dartmouth, for example, have just instituted (in a specified number of courses) a pass-fail alternative to the more rigorously competitive A-F grading system. The interesting aspect of such a change is that there is very little *evidence* to justify it--it might be that some students learn *less* (and are less interested in learning) when their motivation to learn is "intrinsic" (interest in the course material) rather than "extrinsic" (interest in the course grade). One might even argue that incentives (grades) can be utilized as the vehicle for bringing the student to an intrinsic appreciation of learning through a process akin to Allport's notion of Functional Autonomy. The relative advantages and/or disadvantages of competitive grading as a pedagogical technique remain to be defined within the context of empirical research; the point is, such research is vitally needed *now* -- before changes in educational practice take place on the basis of intellectual hunch rather than scientific merit. The major aim of this investigation is to provide some of that necessary research.

This study is concerned with examining performance in course work, and attitudes toward such work, as a function of grading systems that differ in their properties (competitive vs. less competitive) and in such a way provide some empirical information on the relative merits of the two marking systems. It will be emphasized that a pedagogical innovation at Princeton University, allowing the student the choice of taking a course graded in the traditional (A-F) or "new" (pass-fail) manner permits one to study, for the first time, the differential performance by students taking *concurrently* courses with "competitive" and "less competitive" grading systems.

Before presenting the findings of the study it is necessary to describe the Princeton pass-fail system. For many years Princeton undergraduates have been graded on a competitive, highly-delineated numerical marking system employing 17 distinctions from 1+ (best) to 7 (worst). In 1966-67 the faculty modified the standard grading procedure to allow each student the option of taking one course per year on a pass-fail basis (for example, a freshman could elect to take one of his required ten courses pass-fail). To avoid making the pass-fail student a "second-class citizen," instructors were not informed which class members were taking the course pass-fail, and further, were instructed to treat all students alike: pass-fail students were expected to do the same work and be graded in the same way as a student taking the course in the more traditional manner. Only when the numerical grades for all class members were turned into the registrar were the numerical grades of the pass-fail students transformed into a "P" (5- or better) or "F" (6 or 7). In assuring that each student would be treated similarly regardless of his grading choice (pass-fail or 1-7) the University provided the groundwork for a research design that met the four criteria necessary for studying academic performance as a function of different marking systems: (1) students could choose between a competitive and pass-fail grading alternative when signing up for a course; (2) the teacher in the course did not know which students were taking the course pass-fail; (3) the course requirements were the same for each student regardless of the grading system he selected; (4) the teacher marked all students on the basis of the competitive 1-7 system.

This study represents an examination of the pass-fail system as it existed during the 1966-67 academic year (it has undergone modification since then) at Princeton University and reports findings based on questionnaire and performance data from the undergraduate population. Believing that it is just as important to discover what the students think about the pass-fail system as it is to examine their performance within it, the investigation was carried out in two phases: the assessment of attitudes questionnaire phase and the academic performance phase. These two phases will be reported separately.

Phase 1: The Pass-Fail Questionnaire: There are several ways to discover the advantages and limitations of various grading procedures. One of the most direct and relevant ways is to solicit the opinions of students who have experienced them. This method has the obvious advantage of gaining information directly "from the horse's mouth." If the student claims he learns more in a course graded pass-fail than in a similar course taken for A - F credit, one is spared the time necessary to arrive at that answer indirectly (through inference from other data sources). The major drawback of the questionnaire technique is the possible dishonesty of the responding subject who might have a vested interest in the continuation of the pass-fail system. This individual might respond as he thought the University would wish him to respond, rather than as he actually felt. Several steps were taken to encourage honesty in questionnaire responding: (1) instructions emphasized respondent anonymity and the fact that his answers would in no way influence the pass-fail system at Princeton; (2) undergraduates were in charge of questionnaire distribution and utilization, thus underscoring student (not University) concern with the data; (3) one hundred questionnaires were filled out by a random sample of pass-fail students in a personal interview situation conducted by a fellow-student. It was felt that the respondent would be more likely to give a candid, honest view of the pass-fail system in the presence of an undergraduate colleague who assured him the answers he gave would in no way affect the pass-fail system than in a mailed questionnaire which stated the same information in writing.[3]

As a check for respondent honesty on the questionnaires, the results obtained on the one hundred personal interview questionnaires were compared with the findings from the other protocols (906 questionnaires from students who had voluntarily completed and returned them by mail).[4] A second check for honesty was accomplished by giving another group of one hundred pass-fail students a disguised questionnaire designed to solicit pass-fail information while appearing to serve a different purpose. These questionnaires were distributed and collected before the pass-fail questionnaires were delivered. As can be seen by comparing the disguised questionnaire (see Appendix B) with question 35 of the pass-fail questionnaire (see Appendix A) one has a means of estimating a person's judgment of his pass-fail course as a function of varying instructions (realization vs. non-realization that the questions are concerned with attitude toward pass-fail courses). Judging from the: (1) similar response patterns between subjects who filled out the questionnaire through the mail and in a personal interview situation; (2) close correspondence between responses by subjects on the disguised and non-disguised questionnaires; and (3) the rather severe respondent criticisms directed at some aspects of the pass-fail system in both the personal interview and mailed questionnaire sample (comments that would not be made by a student intent on "pleasing the University") it seems relatively safe to claim that the questionnaire results present an accurate, honest picture of the way students feel concerning the pass-fail system at Princeton.

The pass-fail questionnaire (a fifteen minute, 36-item test including true-false, multiple choice and fill-in questions) was designed to assess the respondents: (1) support-non-support for the pass-fail system; (2) reasons for exercising the pass-fail option; (3) perception of the properties of the pass-fail system; and (4) comparison of several facets of the pass-fail vs. numerical grading system. An examination of the students' responses should help answer two questions: (a) just how effective is the pass-fail system as a pedagogical tool?; (b) are the goals of the pass-fail system (the purposes for which it was created) being accomplished?

Opinions of Princeton University undergraduates concerning the continuation of the pass-fail grading system are presented in Table 1.[5] The findings indicate that the pass-fail system has been favorably received by the student body. Only 3 percent of the respondents felt the pass-fail system should be abolished at Princeton; while 48 percent thought more courses should be graded on a pass-fail basis. Further, two-thirds of the respondents were against dropping the pass-fail system in favor of a course reduction program (e.g., instead of five courses, one taken pass-fail; four courses, all for numerical grade).

Table 2 presents findings that allow comparisons between a student's own reasons for taking a course pass-fail and his perceptions of why his classmates and university exercised and implemented the pass-fail option. Almost all the students (87%) thought the University's reason for introducing the pass-fail program was "to encourage course selection on the basis of interest rather than grade considerations"--

TABLE 1

ATTITUDES OF PRINCETON UNIVERSITY UNDERGRADUATES CONCERNING THE CONTINUATION OF THE PASS-FAIL GRADING SYSTEM*

Question	True	False	No Opinion
The pass-fail system should be abolished at Princeton	3	92	5
More courses should be graded on a pass-fail basis	48	31	21
The pass-fail system should be dropped in favor of a course reduction program (e. g., instead of five courses, one taken pass-fail; four courses, all for numerical grade).	22	66	12

*All answers are given in percentages, and are based on the combined random and mailed questionnaire samples (N = 1,006).

yet, only 28 percent took a course pass-fail for that reason, and even fewer (7 %) believed the interest factor influenced their classmates' decision to take a course pass-fail. Instead, 35 percent and 37 percent of the respondents respectively listed "to reduce the tension of, and emphasis on, course grades" and "to provide additional study time for other courses and/or extra-curricular activities" as the reasons they took their course pass-fail. These same reasons accounted for 93 percent of the responses (34 % and 59 % respectively) when the students predicted their classmates' reasons for taking a course pass-fail. Thus, it seems that the majority of students exercise their pass-fail option (and believe their classmates do likewise) to reduce grading tensions and increase additional study time in other courses, rather than on the basis of course interest (the reason, they thought, accounted for the University's implementation of the pass-fail program).

Two of the assumptions made by some critics of numerical (competitive) grading systems is that students will learn more, and their motivation to learn will be higher, in pass-fail compared to numerically graded courses. Information relevant to these beliefs is presented in Table 3. The findings do not support the critics; to the contrary, the findings (from multiple choice questions) indicate that students believe they: (a) learn the same (45 %) or less (41 %) in a pass-fail than comparable numerically graded course; (b) have a decrease in motivation to learn (24 %) or no change in motivation to learn (63 %) when comparing pass-fail with numerically graded courses; and (c) work closer to their capacity (72 %) in a numerically graded course rather than one marked pass-fail.

When the student is asked to make specific percentage comparisons between certain aspects of pass-fail and numerically graded courses, the findings obtained with multiple choice questions are replicated. Utilizing a percentage figure based on the averaged responses of all subjects it was found that a student believes he works to his capacity 67 percent of the time in a numerically graded course, whereas in a pass-fail course he does so only 44 percent of

TABLE 2

WHY TAKE A COURSE PASS-FAIL? A COMPARISON OF A STUDENT'S OWN REASONS WITH THOSE HE ATTRIBUTES TO HIS UNIVERSITY AND FELLOW CLASSMATES*

Reasons	Ss perception of the University's reason for implementing pass-fail	Ss perception of the reason he took a course on a pass-fail basis	Ss perception of the other students' reason for taking a course pass-fail
(1) To reduce the tension of, and emphasis on, course grades.	10	35	34
(2) To provide additional study time for other courses and/or extra-curricular activities.	3	37	59
(3) To encourage course selection on the basis of interest rather than grade considerations.	87	28	7

*All answers are given in percentages, and are based on the combined random and mailed questionnaire samples (N = 1,006).

TABLE 3

ATTITUDES OF PRINCETON UNIVERSITY UNDERGRADUATES CONCERNING HOW MUCH THEY LEARNED AND THE LEVEL OF THEIR MOTIVATION TO LEARN IN PASS-FAIL COMPARED WITH NUMERICALLY GRADED COURSES*

Multiple Choice Questions

% Responding	Question
	As far as the amount of material learned is concerned, I have learned:
0	(a) much more
7	(b) more
45	(c) the same
41	(d) less
7	(e) much less, in my pass-fail than a comparable numerically graded course.
	In my pass-fail course (in comparison to numerically graded courses) I have had:
13	(a) an increase in motivation to learn
24	(b) a decrease in motivation to learn
63	(c) no change in motivation to learn
	I feel that I work closer to my capacity:
72	(a) in a numerically graded course rather than one graded pass-fail
2	(b) in a pass-fail course rather than one graded numerically
26	(c) about equally in both types of courses

Fill-in Questions**

In a numerically graded course I work to my capacity 67 % of the time; in my pass-fail course I do so 44 % of the time.

In a numerically graded course I learn 74 % of the material presented; in my pass-fail course I learn 61 % .

In a numerically graded course I am interested 70 % of the time; in my pass-fail course I am interested 57 % of the time.

True - False Questions

Question	True	False	No Opinion
I was able to learn more in my pass-fail course because it was not graded on a 1-7 basis.	12	69	19
If I was taking my pass-fail course for a normal 1-7 grade, I would be more interested in the course in general.	21	74	5
If I was taking my pass-fail course for a normal 1-7 grade, I would gain more knowledge.	38	48	14
If I was taking my pass-fail course for a normal 1-7 grade, I would gain more intellectual stimulation.	14	73	13

* All answers are given in percentages, and are based on the combined random and mailed questionnaire samples (N = 1,006).

** The figure in the blanks represents the average numerical response for all respondents.

the time. Similar results were obtained when the student compared numerical and pass-fail graded courses in terms of material learned (74 % to 61 %) and interest generated (70% to 57 %) .

An interesting reversal of opinions in favor of the pass-fail grading alternative occurred on three true-false items where students indicated that if they were taking their pass-fail course for a normal 1-7 grade they would not: (a) be more interested in the course in general (74%); (b) gain more knowledge (48%); (c) gain more intellectual stimulation (73%). On a fourth true-false item, however, 69 percent of the respondents answered false to the item: "I was able to learn more in my pass-fail course because it was not graded on a 1-7 basis." A possible explanation for the discrepant results with the true-false items is that they required student estimations based on a hypothetical situation, whereas the multiple choice and fill-in items dealt with actual student behavior in the classroom. Considering the majority of student responses on the majority of questions asked, it seems reasonable to assert that students believe they learn more, work closer to their capacity, and are more motivated to learn in a numerically graded course than one marked pass-fail.

Based on findings reported in Table 3 one might predict that a student's level of participation in a course would be function of whether he was taking the course pass-fail or for numerical grade. The findings reported in Table 4 lend credence to that prediction and indicate that students participate more fully in a course when faced with competitive (1-7)

TABLE 4

ATTITUDES OF PRINCETON UNIVERSITY UNDERGRADUATES CONCERNING COURSE PARTICIPATION IN PASS-FAIL AND NUMERICALLY GRADED CLASSES *

True-False Questions

Question	True	False	No Opinion
If I knew my course was being graded 1-7 rather than pass-fail, I would work harder.	79	18	3
Taking a pass-fail course has allowed me more time to study for my other courses.	74	22	4
As far as I can determine, most of the students have used pass-fail to underemphasize one course while devoting more time to their numerically (1-7) graded courses.	87	8	5
If I had more time I'd work harder in my pass-fail course.	53	42	5
If I was taking my pass-fail course for a normal 1-7 grade,I would do more of the assigned readings.	61	35	4
If I was taking my pass-fail course for a normal 1-7 grade, I would attend more of the lectures and/or precepts.	42	54	4
If I was taking my pass-fail course for a normal 1-7 grade, I would do better work on papers, labs, etc.	50	31	19
If I was taking my pass-fail course for a normal 1-7 grade, I would study harder for course quizzes and exams.	74	22	4

Fill-in Questions **

In a numerically graded (1-7) course I do 80 % of the assigned readings; in my pass-fail course I do 61 %.

In a numerically graded course I attend 85 % of the lectures and/or precepts; in my pass-fail course I attend 74 %.

---

* All answers are given in percentages, and are based on the combined random and mailed questionnaire samples (N = 1,006).

** The figure in the blanks represents the average numerical response for all respondents.

TABLE 5

ATTITUDES OF PRINCETON UNIVERSITY UNDERGRADUATES CONCERNING NUMERICAL (1-7) AND PASS-FAIL GRADING *

True-False Questions

Question	True	False	No Opinion
I have a good idea of how hard I have to work to get a passing grade in my pass-fail course	72	23	5
I have tried to do as well grade-wise in my pass-fail course as in my numerically graded courses	24	73	3
The competition for grades in courses stimulates me to work harder	56	39	5
If I was taking my pass-fail course for a normal 1-7 grade, I would be under more tension due to worry over grades	85	12	3

Fill-in Questions **

Considering the amount of work done in your pass-fail course, what numerical grade would you expect? 3.22.

What kind of grade would you aim for if you were taking the pass-fail course for credit? 2.17.

---

* All answers are given in percentages, and are based on the combined random and mailed questionnaire samples (N = 1,006).

** The figure in the blanks represents the average numerical response for all respondents.

rather than pass-fail grading. A majority of questionnaire respondents agreed that if they were taking their pass-fail course for a normal 1-7 grade they would: (a) work harder; (b) do more of the assigned readings; (c) do better work on papers, labs, etc.; and (d) study harder for course quizzes and exams. They further stated that in a numerically graded course they did 80 percent of the assigned readings and attended 85 percent of the lectures; whereas in their pass-fail course they did only 61 percent of the assigned readings and attended only 74 percent of the lectures. The finding concerning attendance at lectures is somewhat contradictory to the claim of 54 percent of the respondents who answered false to the question: "If I was taking my pass-fail course for a normal 1-7 grade, I would attend more of the lectures and/or precepts."

Why should a student decrease his participation in a pass-fail course? One reason might be an insufficient amount of time to distribute among his classes. In the study, 53 percent of the students claimed they would work harder in their pass-fail course if they had more time. When time is at such a premium, pass-fail courses must compete with numerically graded courses for a student's attention--and under these conditions it is most likely that a student will withdraw time from his pass-fail class and invest it in the "dividend-paying" numerically marked course. In line with these considerations, 74 percent of the students agreed that taking a pass-fail course had allowed them more time to study for their other courses; and an even higher percentage (87 %) believed that most of their classmates had utilized the pass-fail system to underemphasize one course while devoting more time to their numerically (1-7) graded courses. In summarizing the findings reported in Table 4, one can say that a majority of students believed they more actively participated in numerically graded (in comparison to pass-fail) courses, choosing to utilize their limited time allotments in courses where competitive marks were assigned.

Even if one assumes that a student participates more actively in competitively graded courses, concern might still be expressed over the student's reaction to competitive (vs. pass-fail) grading. Findings reported in Table 5 indicate that although 85 percent of the respondents admitted they would be under more tension due to worry over grades in a competitively marked course, this tension might not be useless: 56 percent stated that the competition for grades stimulated them to work harder in their subjects. As might be expected, students believed they worked harder for grades when their courses were competitively marked. Less than one-fourth of the respondents attempted to do as well gradewise in their pass-fail course as in their numerically graded subjects. When asked: "considering the amount of work done in your pass-fail course what numerical grade would you expect?'' the average grade answered was 3. 22. To a second question: ''what kind of grade would you aim for if you were taking the pass-fail course for credit?'' the average grade reply was a healthier 2. 17 (a large difference on the Princeton grading system). Findings from Table 5 indicate that students believe working for competitively assigned 1-7 grades can be tension inducing, but that such grading can also stimulate the student to work harder in the course.

Table 6 presents some general observations by Princeton undergraduates on various aspects of the pass-fail system. It is interesting to note that less than one-third of the respondents found the sub- ect matter of their pass-fail course more difficult than aterial in most of the courses they took for a 1-7 ade. This would seem to indicate that the difficulty the subject matter should not be a factor if it is ound that students receive lower grades in their pass- ail subjects than in courses taken for a numerical (1-7) grade.

Based on undergraduate questionnaire data, one can make the following summary statements (based on majority student opinion) concerning the Princeton pass-fail system: (1) Princeton students are overwhelmingly in favor of retaining the pass-fail grading alternative; (2) the majority of students exercise their pass-fail option (and believe their classmates do likewise) to reduce grading tensions and increase additional study time in other courses, rather than on the basis of interest in the course subject matter; (3) students generally believe they learn more, work closer to their capacity, and are more motivated to learn in a numerically graded course than one marked pass-fail; (4) students believe they participate more fully in a course when faced with competitive (1-7) rather than pass-fail grading; (5) students believe that working for competitively assigned 1-7 grades can be tension inducing, but such grading can also stimulate the student to work harder in his courses.

Phase 2: The Academic Performance Phase:

Trying to discover the properties and assess the worth of the pass-fail system by tapping student opinions is not sufficient; one must also examine the system by studying undergraduate academic performance as a function of differential grading systems. There are two ways in which this task can be accomplished: (1) comparisons of academic performance between

TABLE 6

SOME GENERAL OBSERVATIONS BY PRINCETON UNIVERSITY UNDERGRADUATES CONCERNING THE PASS-FAIL SYSTEM*

Multiple Choice Questions

% Responding	Question
	The best area to take pass-fail courses in is:
15	(a) math and natural sciences
2	(b) engineering
83	(c) liberal arts and social sciences
	The best feature of a pass-fail course is it:
15	(a) lightens your workload
21	(b) can be treated as an ''intellectual experience'' rather than a rigid process of learning and reward
26	(c) allows you to explore areas you might have otherwise hesitated to delve into
12	(d) provides time to learn more in other courses
25	(e) gets rid of grade pressures

True-False Questions

Question	True	False	No Opinion
If I couldn't have taken my pass-fail course on a pass-fail basis, I wouldn't have taken it at all.	26	68	6
Right now, if I had my choice, I'd change my pass-fail course to a 1-7 numerical grade credit.	17	80	3
The subject matter of my pass-fail course was more difficult than material in most of the courses I take for a (1-7) grade.	31	65	4

---

* All answers are given in percentages, and are based on the combined random and mailed questionnaire samples (N=1,006).

TABLE 7

MEANS AND STANDARD DEVIATIONS FOR THE TOTAL SUBJECT SAMPLE

Variable	$\bar{X}$	SD	N
Number of courses taken during the Academic Year	8.94	1.45	2,986
Grade Point Average for the Academic Year	2.43	.69	2,975
SAT Verbal Score	646	67.2	2,925
SAT Math Score	691	65.8	2,916

individuals who chose to exercise their pass-fail option and those who did not; (2) comparisons of academic performance by pass-fail students in their pass-fail and competitively graded courses.

For the purpose of baseline comparisons, Table 7 presents the means and standard deviations for relevant academic variables based on the total subject sample. Although most of these students (N = 2,475+) chose to exercise their pass-fail option, a small number (N = 425+) did not. Table 8 presents a comparison of means and standard deviations for these pass-fail and non-pass-fail students on relevant academic measures. The results point to an important finding: there is no significant difference between yearly grade point averages for pass-fail and non-pass-fail students (F = 2.07, NS). This was determined by calculating a yearly grade point average for non-pass-fail students (based on their grades in all courses for the academic year) and comparing it with the yearly grade point average for pass-fail students (calculated on the basis of performance in all courses except the pass-fail class). If, as the questionnaire data suggest, pass-fail students gain additional time for study in other courses by exercising their pass-fail option, then the above results indicate that this extra time did not lead to superior performance over those students who had no such extra time.

A second finding of importance supports the expectation ($t = Pvt < .00001$) that students will receive better grades in courses graded competitively than in those marked on a pass-fail basis. This was determined by recording each pass-fail student's numerical grade from instructors' grade sheets before it was changed to a "pass" (5-or better) or "fail" (6 or 7) by the registrar's office. This grade was then utilized for comparison with numerical grades achieved by the student in his competitively graded courses. The large differences in academic performance as a function of differential grading systems can be noted by comparing the mean pass-fail grade (3.25) for all pass-fail students with their mean grade point average for the year (2.42).[8] This finding supports questionnaire data where respondents indicated they worked harder for grades in competitively marked (1-7) courses.

Based on academic performance data, one can make the following summary statements concerning the Princeton pass-fail system: (1) students do equally well on their yearly grade point average irregardless of whether or not they exercise their pass-fail option; (2) students get significantly better grades in their competitively graded courses than in their pass-fail subject.

TABLE 8

A COMPARISON OF MEANS AND STANDARD DEVIATIONS FOR PASS-FAIL AND NON-PASS-FAIL STUDENTS

	Pass-Fail Students			Non-Pass-Fail Students		
Variable	$\bar{X}$	SD	N	$\bar{X}$	SD	N
Number of Courses taken during the academic year	8.95	1.44	2,535	8.92	1.51	448
Pass-Fail grade	3.25	1.05	2,486	-	-	-
Grade Point Average for the academic year	2.42	.67	2,525	2.47	.75	447
SAT Verbal Score	644	66.7	2,490	655	69.4	432
SAT Math Score	690	66.2	2,484	695	63.2	429

DISCUSSION

The purpose of this study was to examine student performance in course work, and attitudes toward such work, as a function of grading systems that differed in their properties (competitive vs. less competitive) and in such a way provide some empirical information on the relative merits of the two marking systems. The results of this investigation indicate that the pass-fail grading alternative is not the panacea that some educators envisioned. This is not to imply that a pass-fail system is worthless, only that findings reported here underscore the need for further study and reexamination of the pass-fail system before it becomes a permanent grading procedure in our Universities. The urgency of this need is indicated in the rather unexpected responses of many students who often believed they learned more, worked closer to their capacity, were more motivated to learn and more actively participated in numerically graded courses than ones marked pass-fail. These beliefs were reflected in actual performance: students received significantly better grades in their competitively graded courses than in their pass-fail subjects. If, indeed, students believe they learn more in competitively graded courses, one might ask why a pass-fail grading alternative should exist at all. Several answers could be given to this question (based on experimental findings) including: (1) a willingness on the part of some students to explore courses they would ignore if faced with a 1-7 grade; (2) a reduction in tension due to the elimination of competitive grading. Findings from the study also suggest that the effectiveness of the pass-fail system is undermined when a student's course load creates time pressures in his study schedule. When time is at such a premium, pass-fail courses must compete with numerically graded courses for a student's attention--and under these conditions it is most likely that a student will withdraw time from his pass-fail class and invest it in the "dividend-paying" numerically marked course. This is especially true at a school like Princeton where the undergraduates are accustomed to competitive grading and also have a high level of achievement motivation. In an academic system where all courses are marked pass-fail or where students have sufficient time to devote themselves in their pass-fail studies a pass-fail grading alternative will probably come much closer to fulfilling the goals for which it was implemented.

FOOTNOTES

1. This research was supported by an NSF Institutional Grant awarded to the investigator by Princeton University.

2. The author gratefully acknowledges the generous assistance of the following individuals: Gene Bruskin, John Hayes, Jim Kelly, and Tom Madeira for their aid in the questionnaire phase of the investigation; Sandra Bull for her data recording and processing; Thomas Coffman, Bertram Koslin and Richard Pargament for their assistance in data analysis and, finally, to the Princeton student body for their aid in the project.

3. In addition to providing a check for respondent honesty, the personal interview questionnaire subjects served as a random sample of the pass-fail student population. Such a random sample was carefully selected and tested in anticipation that the mailed-questionnaire respondents might not represent an accurate cross-section of all pass-fail students.

4. This figure represents just under a 50 percent return on questionnaires mailed out.

5. Questions were not given to students in the order appearing in the Table. See Appendix A for actual order of presentation.

6. Note how closely the actual mean pass-fail grade (3.25) corresponds to the students mean estimation (Table 5) of what their pass-fail grade will be (3.22).

APPENDIX A

PASS-FAIL QUESTIONNAIRE

Section 1.

True-False. In this section please circle a "T" if you think the item is True; circle "F" if you think it is False; circle "0" if you have no Opinion concerning the item. Please answer these questions based on your experience with this semester's Pass-Fail course!

T True F False 0 No Opinion

T F 0 (1) As far as I can determine, most of the students have used pass-fail to under-emphasize one course while devoting more time to their numerically (1-6) graded courses.

T F 0 (2) I have a good idea of how hard I have to work to get a passing grade in my pass-fail course.

T F 0 (3) Taking a pass-fail course has allowed me more time to study for my other courses.

T F 0 (4) The pass-fail system should be dropped in favor of a course reduction program (e.g., instead of five courses, one taken pass-fail; four courses, all for numerical grade).

T F 0 (5) Right now, if I had my choice, I'd change my pass-fail course to 1-6 numerical grade credit.

T F 0 (6) If I couldn't have taken my pass-fail course on a pass-fail basis I wouldn't have taken it at all.

T F 0 (7) More courses should be graded on a pass-fail basis.

T F 0 (8) The pass-fail system should be abolished at Princeton.

T F 0 (9) I devote as much time to my pass-fail course now as I thought I would when I signed up for the course.

T F 0 (10) The competition for grades in courses stimulates me to work harder.

T F 0 (11) The subject matter of my pass-fail course was more difficult than material in most of the courses I take for a (1-6) grade.

T F 0 (12) If I had more time I'd work harder in my pass-fail course.

T F 0 (13) If I knew my course was being graded 1-6 rather than pass-fail, I would work harder.

T F 0 (14) I didn't explore my pass-fail coursework fully because I knew it wouldn't affect my grade.

T F 0 (15) I have tried to do as well gradewise in my pass-fail course as in my numerically graded courses.

T F 0 (16) I was able to learn more in my pass-fail course because it was not graded on a 1-6 basis.

(17-24) If I was taking my pass-fail course for a normal 1-6 grade I would:

T F 0 (17) Do more of the assigned readings.

T F 0 (18) Attend more of the lectures and/or precepts

T F 0 (19) Study harder for course quizzes and exams

T F 0 (20) Do better work on papers, labs , etc.

T F 0 (21) Be more interested in the course in general

T F 0 (22) Be under more tension due to worry over grades

T F 0 (23) Gain more knowledge

T F 0 (24) Gain more intellectual stimulation

Section 2.

Multiple choice. In this section please place the number of the alternative that best expresses your sentiment concerning the question to the left of that question number. Choose only one alternative per question. Please answer these questions based on your experience with this semester's Pass-Fail course!

_______ (25) The best feature of a pass-fail course is it:
(1) Lightens your workload.
(2) Can be treated as an "intellectual experience" rather than a rigid process of learning and reward.
(3) Allows you to explore areas you might have otherwise hesitated to delve into.
(4) Provides time to learn more in other courses.
(5) Gets rid of grade pressures.

_______ (26) I am learning: (1) much more (2) more (3) the same (4) less (5) much less, than I expected from my pass-fail course.

_______ (27) The worst feature of a pass-fail course is:
(1) It is often wasted since I tend to neglect it when time gets short.
(2) I still worry about passing it when I don't do all the necessary work.
(3) It's an option I wouldn't exercise except that everyone else does and if I didn't they would gain an unfair time advantage over me.
(4) The fear of getting a "good grade" for nothing.
(5) None of the above.

_______ (28) I feel that I work closer to my capacity:
(1) In a numerically graded course rather than one graded pass-fail
(2) In a pass-fail course rather than one graded numerically
(3) About equally in both types of courses

_______ (29) In my pass-fail course (in comparison to numerically graded courses) I have had:
(1) an increase in motivation to learn
(2) a decrease in motivation to learn
(3) no change in motivation to learn

_______ (30) As far as the amount of material learned is concerned, I have learned:
(1) much more (2) more (3) the same (4) less (5) much less, in my pass-fail than a comparable numerically graded course.

_______ (31) The best area to take pass-fail courses in is:
(1) math and natural sciences
(2) engineering
(3) liberal arts and social sciences

_______ (32) Which of the following factors was most important in your choice of a pass-fail course ?
(1) the difficulty of getting a good numerical grade
(2) the subject matter of the course was unfamiliar and/or more difficult than in other courses taken for numerical credit
(3) the subject matter of the course was familiar and/or less difficult than in other courses taken for numerical credit
(4) specific requirements of the course: (e.g., paper, exams, etc.).
(5) none of the above

Section 3.

Miscellaneous. Please fill in each question according to directions. Please answer these questions based on your experience with this semester's Pass-Fail course.

(33) Considering the amount of work done in your pass-fail course, what numerical grade would you expect ? __________ .

What kind of grade would you aim for if you were taking the pass-fail course for credit ? _______.

---

(34) Below are three possible goals that might be achieved in giving pass-fail courses at Princeton (you can use a goal more than once in answering the questions):
(1) To reduce the tension of, and emphasis on, course grades.
(2) To provide additional study time for other courses and/or extra-curricular activities.
(3) To encourage course selection on the basis of interest rather than grade considerations.

_______ In the space to the left write the number of the goal that you think best describes the University's reason for implementing the pass-fail system.

_______ In the space to the left write in the number of the goal that you think best describes your reason for taking a course on a pass-fail basis.

_______ In the space to the left write in the number of the goal that you think best describes the reason why most students at Princeton take a course pass-fail.

---

(35) In the blank that follows write down the number of courses (pass-fail included) you took this semester _______. Then please circle the numbers below that best indicate the way you would compare your pass-fail course with other courses taken during this spring semester.

(a) Of my courses this semester, my pass-fail course has been the 1st (most), 2nd, 3rd, 4th, 5th, 6th (least) intellectually stimulating (interesting) course. (NOTE: If you took four courses and the pass-fail was the most intellectually stimulating you should have circled a "1st" above; if it was the second most stimulating course, a "2nd" above; if it was least stimulating of the 4 courses a "4th" above, etc. If you had taken five courses instead of four and the pass-fail was least stimulating of these five, you should have circled a "5th" above.)

(b) Of my courses this semester, I have learned the 1st (most), 2nd, 3rd, 4th, 5th, 6th (least) amount of information in my pass-fail course.

(c) Of my courses this semester, the subject matter of my pass-fail course has been the 1st (most), 2nd, 3rd, 4th, 5th, 6th (least) difficult.

(d) Of my courses this semester, I studied 1st (most), 2nd, 3rd, 4th, 5th, 6th, (least) in my pass-fail course.

(e) Of my courses this semester, my motivation to learn was 1st (most), 2nd, 3rd, 4th, 5th, 6th (least) in my pass-fail course.

---

(36) Utilize the percentages listed below in answering this final set of questions about your spring semester pass-fail course. In each blank place the percentage figure that most closely represents the true amount of effort you invested in a specified activity.

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

(a) In a numerically graded (1-6) course I do ______% of the assigned readings; in my pass-fail course I do ________%

(b) In a numerically graded course I attend _______% of the lectures and/or precepts; in my pass-fail course I attend ________%.

(c) In a numerically graded course I work to my capacity________% of the time; in my pass-fail course I do so ______% of the time.

(d) In a numerically graded course I am interested______% of the time; in my pass-fail course I am interested________% of the time.

(e) In a numerically graded course I learn_______% of the material presented; in my pass-fail course I learn ________%.

Thank you for your cooperation.

APPENDIX B

COURSE COMPARISON QUESTIONNAIRE

INSTRUCTIONS:

We are interested in finding out what you think about your current spring semester courses. Here is a list of those courses. We would like you to compare these courses by rating them from "most" to "least" on certain educational attributes. For example, we want to know how you rate your courses according to their intellectual stimulation. Thus, for this aspect of your courses, you would list first that course which was most intellectually stimulating of those you took this semester; second, the course that was next most intellectually stimulating, and so on down to the course that was least intellectually stimulating. Follow this procedure for the ten items below. Include all your courses in each rating and please do not permit any "ties" between courses in your ratings.

(1) Please list below the courses you took in order of how intellectually stimulating (interesting) they were:

Most intellectually stimulating

Least intellectually stimulating

(2) Please list below the courses you took in order of the amount of information learned in them:

Most information learned

Least information learned

(3) Please list below the courses you took in order of how difficult the subject matter was in them:

Most difficult subject matter

Least difficult subject matter

(4) Please list below the courses you took in order of how hard you studied in them:

Studied Most

Studied Least

(5) Please list below the courses you took in order of how motivated you were to learn in them:

Most motivated to learn

Least motivated to learn

(6) Please list below the courses you took in order of how much grade pressure there was:

Most grade pressure

Least grade pressure

(7) Please list below the courses you took in order of how many lectures and/or precepts you attended:

Most lectures/precepts attended

Least lectures/precepts attended

(8) Please list below the courses you took in order of how much of the assigned reading you did:

Most assigned reading done

Least assigned reading done

(9) Please list below the courses you took in order of how close you came to working up to your capacity in them:

Worked closest to capacity

Worked least closest to capacity

(10) Please list below the courses you took in order of how hard you worked for grades in them:

Tried hardest for grade

Tried least for grade

ders
**DEMBAR
EDUCATIONAL
RESEARCH
SERVICES, INC.**

POST OFFICE BOX 1148 · MADISON, WISCONSIN 53701

TITLES FROM DERS

SCHOLARS' GUIDE TO JOURNALS OF EDUCATION

And Educational Psychology

By L. Joseph Lins and Robert A. Rees. 1965

160 pages, $3.95

If you must publish or perish, here is an annotated index to 135 leading educational journals, with answers to such key questions as how to write, whom to write for, how to submit your manuscript, and so on.

"Invaluable information."—THE SCIENCE TEACHER.

YEAR-ROUND EDUCATION

**Its Problems and Prospects
from Kindergarten to College**

By Clarence A. Schoenfeld and Neil Schmitz. With a foreword by John R. Little. Second Edition, 1966.

144 pages, $3

What is the truth about the all-year calendar? Is year-round education panacea or pitfall? Here is a unique assessment of a national issue that focuses our attention on demonstrable facts rather than on popular fancy. Chapters deal with the all-year elementary and secondary school and with the year-round college and university.

"An outstanding piece of work."—Dr. FRANK BURRIN, Purdue University.

** The figure in the blanks rep-

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

PERCEPTUAL-COGNITIVE STYLE AS RELATED TO SELF-EVALUATION AND SUPERVISOR RATING BY STUDENT TEACHERS

DONALD MUSELLA

Ontario Institute for Studies in Education

ABSTRACT

Research in the area of social perception has pointed to (1) the relationship between certain personal characteristics and the perception and cognition of others, (2) the effects of certain personal characteristics on judgments of others, and (3) the influence of dogmatism on the perceptual-cognitive style affecting judgments of others. An earlier study by this researcher provided evidence supporting the effect of dogmatism on the description and rating of elementary school teachers by elementary school principals. This study represents an effort to provide additional information related to (1) the relationship between dogmatism and self-evaluation in teaching, and (2) the relationship between dogmatism and the evaluation of supervising personnel. A review of literature and of previous work in this area by the researcher produced hypotheses related to differences between openminded and closedminded student teachers and their rating and description of themselves and of their supervising teachers. The results support the hypotheses in this study, as well as the results of the previous study. These results indicate that the rating of one's teaching, which may be considered a manifestation of the evaluation of self in total, and the rating of one's superordinates are, in some respects, a function of the perceptual-cognitive style of the rater.

RESEARCH IN the area of social perception has pointed to the relationship between certain personal characteristics and the perception and cognition of others (1, 2, 3, 4, 7, 8, 10). Several studies have identified the effects of certain personal characteristics on judgments of others (6, 9, 14, 21, 22, 23, 27, 28). Other studies have dealt more specifically with the influence of dogmatism on the perceptual-cognitive style affecting judgments of others (5, 12, 13, 15, 16, 19, 21, 23, 25, 29).

An earlier study by this researcher provided evidence supporting the effect of dogmatism on the description and rating of elementary school teachers by elementary school principals (16).

The present study represents an effort to provide additional information related to 1) the relationship between dogmatism and self-evaluation in teaching, and 2) the relationship between dogmatism and the evaluation of supervising personnel.

While the results of the earlier study were clear and unambigious, the limitation of sample (elementary teachers and administrators) raises questions about the generality of the findings. The present study constitutes a partial replication, aimed primarily at determining whether the interrelationships emerge under different conditions, with different populations, and with an additional supportive measurement device.

HYPOTHESES

A review of literature and of the previous work in this area by this researcher produced the following hypotheses:

(1) closed-minded student teachers would tend to rate themselves higher in teaching than would open-minded student teachers;

(2) closed-minded student teachers would tend to rate supervising teachers lower in teaching than would open-minded student teachers;

(3) closed-minded student teachers would tend to describe themselves in more positive terms than would open-minded student teachers;

(4) closed-minded student teachers would tend to display less differentiation and variability in their descriptions of teaching than would open-minded student teachers;

(5) both groups of student teachers would tend to select identical characteristics considered essential for effective teaching.

METHODS

Sample

The sample was composed of 128 student teachers enrolled at State University College at Cortland, New York. Each student teacher was assigned for one 8-week period to one of sixteen supervising teachers in the Campus-Laboratory School. In most cases, the student teachers were selected by the supervising teachers on the basis of the initial teaching experience provided for each college student during the previous year. These students were considered, by the supervising teachers, to be potentially above-average student teachers.

The total sample consisted of thirty-two male and ninety-six female students who were enrolled in the elementary education program during their junior year of college. The supervising teacher group was comprised of five male and eleven female teachers of grades nursery through 6. All supervising teachers were full-time professional staff assigned to the laboratory school.

From this sample, two groups were identified: the upper 25 percent scoring high in dogmatism (closed-minded student teacher -- CST), and the lower 25 percent scoring low in dogmatism (open-minded student teacher -- OST).

Instruments

Three instruments were used in this study: (1) Dogmatism Scale (D scale--Form E); (2) Teacher Personal Characteristic Inventory (TPCI); and (3) Teaching Evaluation Form (TEF).

The D Scale was developed by Rokeach and his associates (25). As a partial validation of both the theory and scale, the investigators identified subjects who scored high on the D Scale (closed-minded subjects) and those scoring low on the scale (open-minded subjects) and subjected them to various tests, problems, and other analyses. In the samples used, the reliabilities ranged from .68 to .93.

The TPCI, developed by this researcher (16) is a modification of the rating scale developed by Ryans (26); it is patterned after the semantic-differential scale of Osgood, and others (18). The TPCI consists of twenty-three pairs of polar terms, arrived at by means of the critical incident technique, which represent behaviors of the most effective and the least effective teachers. These characteristics were derived from over 500 critical incidents submitted by practicing administrators, teachers, practice-teachers, and college teachers.

The TEF, developed by this researcher (17), is a modification of the rating scale developed by Riley, and others (20). This scale consists of fifteen items to be responded to by a numerical expression of one of 6 degrees of effectiveness. In addition, most effective and least effective teaching qualities are selected as descriptive of the teacher being rated.

TABLE 1

COMPARISON OF SCORES BETWEEN MALE AND FEMALE STUDENT TEACHERS

	Male	Female	Total
Number	32	96	128
D Scale			
Range	81-189	90-195	81-195
Mean	141.7	140.8	144.4
S.D.	24.1	22.5	21.7
not significant			
TPCI			
Range	82-130	88-158	82-158
Mean	122.4	120.3	124.2
S.D.	23.6	21.9	22.1
not significant			
TEF_s			
Range	29-86	31-82	29-86
Mean	62.7	60.0	61.1
S.D.	8.1	9.5	9.2
not significant			
TEF_{sp}			
Range	41-80	45-90	41-90
Mean	70.5	72.9	72.7
S.D.	8.4	9.2	9.3
not significant			
AGP*			
Range	2.0-3.1	2.0-3.5	2.0-3.5
Mean	2.9	3.1	2.8
S.D.	.5	.6	.4
not significant			

*Accumulative Average Grade Point. Students with less than 2.0 at the end of the sophomore year were not permitted to student teach (department policy).

Procedure

Student teachers were assigned to the Campus-Laboratory School for a period of 8-weeks. The specific assignment was based on the interests of the college students (their requests) and the approval of the supervising teacher to whom they were assigned.

During the first week of the student teaching assignment, each student teacher completed the D Scale. After the conclusion of this 8-week teaching experience, and after the supervisor had submitted final grades, each student teacher was asked to 1) rate the supervising teacher on teaching (not supervision) using the TEF, 2) rate himself/herself on teaching on the identical scale--TEF, 3) describe himself/herself on the TPCI, and 4) select, from the list of positive traits on the TPCI, the five characteristics most essential to effective teaching.

The controls inherent within the study were as follows:

(1) all student teachers spent a relatively equal amount of time in immediate contact with each supervisor; the situation was such that both the supervisor and student taught the same class every day; each had ample opportunity to observe each other's teaching in the self-contained classroom setting;

(2) all student teachers had similar formal educational experiences; all were elementary education majors in their 3rd year at the same college;

(3) all assignments were based on the choice of both the student teacher and the supervising teacher;

(4) none of the student teachers had any previous formal teaching experience; and

(5) there were no significant differences between the male and female student teachers on the accumulative average grade point at the end of the previous academic year. (See Table 1.)

TABLE 2

COMPARISON OF TEF_s SCORES BETWEEN CST AND OST

	CST	OST
Range	39-86	31-82
Mean	71.3	62.4
S.D.	7.6	8.2
n	32	32

t = 1.99; significant at .05

COMPARISON OF TEF_{sp} SCORES BETWEEN CST AND OST

	CST	OST
Range	41-82	51-86
Mean	62.3	79.5
S.D.	8.4	8.9
n	32	32

t = 2.23; significant at .05

TABLE 3

COMPARISON OF TPCI SCORES BETWEEN CST AND OST

	CST	OST
Range	105-158	91-158
Mean	140.4	129.6
S.D.	15.0	15.3
d.f.	31	31

t = 3.87; significant at .01 level

Item Scores

	CST	OST
Range	3 - 6	2 - 6
Mean	5.71	6.29
S.D.	.27	1.11
Var.	.081	1.152
d.f.	31	31

F = 10.39; significant at $.01^+$ level

RESULTS

A summary of data received from the four scales--D Scale, TPCI, TEF_s (self), and TEF_{sp} (supervising teacher), indicate no significant differences between male and female students on any of the scales. (See Table 1.) Therefore, the total student teacher group can be considered a relatively homogeneous sample in terms of sex differences related to behavior on the scales and to success in college.

The statistical results, as provided in Tables 2, 3, and 4, are summarized as follows:

(1) closed-minded student teachers (CST) rated selves higher on teaching effectiveness than did open-minded student teachers (OST) (Table 2);

(2) CST rated supervisors lower on teaching effectiveness than did OST (Table 2);

(3) CST described themselves in more positive terms (high rating scores) than did OST (Table 3);

(4) CST displayed less variability in their descriptions of self than did OST (Table 3);

(5) CST and OST did not differ significantly in the selection of five characteristics most essential to effective teaching; CST selected (a) stimulating, (b) responsible, (c) systematic, (d) confident, and (e) fair; OST selected (a) responsible, (b) confident, (c) stimulating, (d) adaptable, and (e) understanding. The correlation between groups was a highly positive one (.91). (See Table 4.)

TABLE 4

COMPARISON OF CHARACTERISTICS SELECTED BY CST AND OST

Characteristics	No. CST	No. OST
stimulating	6	5
responsible	5	7
systematic	5	1
confident	4	5
fair	4	2
democratic	2	1
understanding	2	4
active	1	1
adaptable	1	5
optimistic	1	0
original	1	1
permissive	1	0
responsive	1	0
alert	0	0
altruistic	0	0
attractive	0	0
broad	0	0
integrated	0	0
kindly	0	0
poised	0	0
positive	0	0
relative	0	0
steady	0	0

r = .91 (Spearman rho)

DISCUSSION, CONCLUSIONS, AND IMPLICATIONS

The results support the hypotheses in this study, as well as the results of the previous study by this researcher (16). These results indicate that the rating of one's teaching, which may be considered a manifestation of the evaluation of self in total, and rating of one's superordinates is, in some respects, a function of the perceptual-cognitive style of the rater. It is indicated further that rating results could be reliably predicted from the theoretical development of Rokeach's investigations on individual personality. Not only did the findings lend support to the Rokeach theory, but also they suggested that personal characteristics of the rater are related to the rating of teaching, both of one's self and of one's supervisors. On the basis of the evidence cited, real differences are discerned between predominantly open and predominantly closed student teachers in their rating and discription of teaching effectiveness. Of course, it is possible that the differences could be the result of differences in individuals, especially since the subjects had no previous teaching experience upon which to base judgments. However, similar results were found with experienced principals and teachers (16). Furthermore, it was found that differences existed between open and closed student teachers in the description of self on the TPCI. Closed subjects displayed less differentiation and variability in their descriptions. They tended to see themselves in completely positive terms, whereas the open subjects pointed to both positive and negative characteristics in their descriptions of self.

No significant differences were found in the selection of characteristics considered to be most necessary for effective teaching. In supporting this finding, one might speculate that the exposure to the "desired traits" provided students in the literature and in the language of the college classroom provides a ready list that can be used to characterize the ideal teacher. However, this latter point is one worthy of further study.

The implications of this study bear direct relevance to existing theories of perceptual-cognitive style and to theories which employ a social-psychological orientation to interpersonal relations (especially interpersonal perception) in a social and organizational context. The importance of personal characteristics becomes apparent in the examination of other forms of decision-making at all levels of formal and informal superordinate-subordinate role relationships.

The limitations of this study are readily apparent in terms of population size, population characteristics, measurement devices, and the quantity and quality of interpersonal contact. Replication of results must emphasize considerable control on the above-mentioned variables, especially the interpersonal contact. In addition, improvement of the measures of teacher rating is essential in order to identify more specific aspects of teaching. The measurement devices employed in this study were restricted to "global" definitions of teaching.

REFERENCES

1. Bieri, J., "Cognitive Complexity-simplicity and Predictive Behavior," Journal of Abnormal Social Psychology, 51:263-268, 1955.
2. Brown, Alan F., "A Perceptual Taxonomy of the Effective-rated Teacher," Journal of Experimental Education, 35: 1-11, 1966.
3. Bruner, J. and Tagiuri, R., "The Perception of People," in G. Lindzey (Ed.), Handbook of Social Psychology, 1954, Cambridge, Massachusetts, Addison-Wesley, pp. 634-654.

4. Cantril, H., "Perception and Interpersonal Relations," American Journal of Psychiatry, 114: 119-126, 1957.

5. Costin, F., "Dogmatism and Learning: a Follow-up of Contradictory Findings," Journal of Educational Research, 59:186-188, 1965.

6. Dailey, C. A., "The Effects of Premature Conclusions Upon the Acquisition of Understanding of a Person," Journal of Psychology, 33:133-152, 1952.

7. Erickson, E. A., "A Study of the Relationship Between Personality Structure and Perception," unpublished doctoral thesis, Colorado State College, 1962.

8. Frenkel-Brunswik, E., "Intolerance of Ambiguity as an Emotional and Perceptual Personality Variable," Journal of Personality, 18:108-143, 1949.

9. Hastorf, A. H., Richardson, S. A., and Dornbusch, S. M., "The Problem of Relevance in the Study of Person Perception," in R. Tagiuri and L. Petrullo (Eds.), Person Perception and Interpersonal Behavior, Stanford University Press, Stanford, California; pp. 54-62.

10. Heider, F., The Psychology of Interpersonal Relations, John Wiley and Sons, Inc., New York; 1958.

11. Hellkamp, D. J. and Marr, J., "Dogmatism and Field Dependency," Perceptual and Motor Skills, 20:1046-48, 1965.

12. Jones, E. E., "Authoritarianism as a Determinant of First-impression Formation," Journal of Personality, 23:107-127, 1954.

13. Kemp, C. G., "Improvement of Critical Thinking in Relation to Open-closed Belief Systems," Journal of Experimental Education, 31:321-323, 1963.

14. Luchins, A. S., "On Agreement with Another's Judgment," Journal of Abnormal Social Psychology, 39:97-11, 1944.

15. Miller, N., "Involvement and Dogmatism as Inhibitors of Attitude Change," Journal of Experimental and Social Psychology, 121-32, 1965.

16. Musella, D. F., "Open-closed Mindedness As Related to the Rating of Teachers by Elementary School Principals," Journal of Experimental Education, 35:75-79, 1967.

17. Musella, D. F. and Rusch, R., "Student Opinion on College Teaching," Improving College and University Teaching, Spring: 132-140, 1968.

18. Osgood, E. E., Suci, G. J., and Tannenbaum, P. H., "The Measurement of Meaning," University of Illinois Press, Urbana, 1957.

19. Plant, W. J., Telford, C. W. and Thomas, J. A., "Some Personality Differences Between Dogmatic and Non-dogmatic Groups," Journal of Social Psychology, 67: 67-75, 1965.

20. Riley, J. W., Ryan, B. F., and Lifshitz, M., "The Student Looks at his Teacher," Rutgers University Press, New Brunswick, New Jersey, 1950.

21. Robbins, L. Y., "The Dogmatism of Teachers," Journal of Teacher Education, 17:47-49, 1966.

22. Rokeach, M., "Generalized Mental Rigidity as a Factor in Ethnocentrism," Journal of Abnormal Social Psychology, 43: 254-278, 1948.

23. Rokeach, M., "Rigidity and Ethnocentrism: a Rejoiner," Journal of Personality, 17:467-474, 1949.

24. Rokeach, M., "Narrow-mindedness and Ethnocentrism," American Psychologist, 5:308, 1950.

25. Rokeach, M., The Open and Closed Mind, Basic Books, Inc., New York, 1960.

26. Ryans, D., "Characteristics of Teachers: Their Descriptions, Comparisons, and Appraisals," American Council on Education, Washington, D. C. 1960.

27. Sorkin, F. F., "Forming Impressions of Others," Journal of Personality, 22: 118-127, 1953.

28. Taft, R., "The Ability to Judge People," Psychological Bulletin, 52: 1-23, 1955.

29. Vidulich, R., and Kaiman, I. P., "The Effects of Information Source Status and Dogmatism on Conformity Behavior," Journal of Abnormal Social Psychology, 63: 639-642, 1961.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

LONGITUDINAL RELATIONSHIPS BETWEEN COGNITIVE AND AFFECTIVE LEARNING OUTCOMES

CHARLES O. NEIDT
Colorado State University
and
DALVA E. HEDLUND
Cornell University

ABSTRACT

The present study was designed to investigate the longitudinal covariation of attitude toward subject matter and achievement during a course in general psychology. Five measures of attitudes toward psychology and five measures of achievement were obtained at intervals of 20 percent through the course. Subjects included 866 freshmen and sophomores at Colorado State University and the University of Missouri. The data indicated a consistent improvement in achievement throughout the course but relatively little change in attitudes. It was concluded that the cognitive and affective outcomes were independent of each other.

PREVIOUS STUDIES have shown attitude toward a subject may be related to performance in a course of instruction (1, 2), and that attitude toward the instructional situation itself may be related to final course achievement (4). It was pointed out in the Taxonomy of Educational Objectives: Affective Domain (3) that the implicit assumption behind most teaching efforts is that if the cognitive outcomes of a learning experience achieved there will automatically be a corresponding development of appropriate affective behaviors. However, there is little, if any, empirical data on the relationship over time of affective and cognitive learning outcomes.

The present study was designed to investigate the longitudinal covariation of attitude toward the subject matter (the affective outcome) and achievement (the cognitive outcome) during a course of instruction.

METHOD

Situation and Subject

The longitudinal design of study required repeated measures of attitude toward the subject matter (affective outcomes) and achievement (cognitive outcome), both obtained at the same times during a course of instruction. Introductory Psychology was chosen as the subject since large numbers of students were involved and the cooperation of instructors could most easily be obtained. Three separate courses of instruction were involved: the Fall, 1966, Introductory Psychology class at Colorado State University (CSU) (N=384); the Winter, 1967, Introductory Psychology class at CSU (N=297); and the Spring Semester Introductory Psychology class at the University of Missouri (N=185). All three classes were taught by closed-circuit television lectures, and all classes met in sections of approximately thirty to forty students taught by graduate assistants. The lecture series presented to both CSU classes was on tape and thus virtually identical between classes, except that the Fall class had a few extra tapes.

Attitude Toward Psychology Scale

To develop an attitude toward psychology scale, a taxonomy of areas in which attitudes were probably instructional objectives was developed from an examination of several widely used psychology textbooks. Statements were then drawn from CSU text and taped lectures to conform to the attitude taxonomy. A total of ninety-seven statements were constructed and edited for clarity and brevity. These statements were then submitted to eight judges, three regular faculty members, and five graduate assistants who had experience with Introductory Psychology. The instructors judged whether they attempted to teach an attitude, to discourage the attitude, or were neutral toward it in class. Only items on which all judges agreed perfectly about encouraging or discouraging were considered for inclusion in the final attitude scale. Ten items representing the attitude taxonomy were chosen for the final attitude scale. Only ten items were chosen to minimize administration time while assuring the possibility of obtaining satisfactory reliability. The items were set in a Likert 5-point constant response scale which was weighted

0-4. Scores on the attitude scale were the sums of the weighted responses to the ten items.

It was decided to measure attitude toward psychology and achievement after each successive 20 percent of a course of instruction. To accomplish the repeated measurements, five equivalent forms of the attitude scale were developed by changing unimportant words in attitude statements and randomly rearranging items within the scale. The discrimination power of each item was checked on a random sample of one hundred subjects drawn from the CSU Fall class and divided into upper and lower 50 percent groups on the basis of total attitude score. Split-half reliability coefficient, corrected by Spearman-Brown prophecy formula, were calculated for all five forms of the attitude scale from from data on this same one hundred Ss. The corrected coefficients ranged from .35 to .83, with most at about .60. Correlation coefficients between adjacent administrations of the attitude scales were also calculated in all three classes, yielding an estimate of test-retest reliability over a period of 2 to 3 weeks between different forms of the attitude scale. The fifteen test-retest reliability coefficients ranged from .56 to .83, with all but three coefficients above .66.

Achievement Tests

Five successive achievement measures were also necessary to detect changes in achievement during a course of instruction. A pool of approximately 500 objective achievement items was constructed from an analysis of the CSU textbook and TV lectures. Items in this pool were administered to seventy-seven students enrolled in a summer session Introductory Psychology course. Items which successfully differentiated high from low achievers in this pretest series were maintained in the item pool. Since the primary purpose of a repeated achievement measure was to reliably differentiate students on the basis of class achievement or cognitive learning, a traditional learning curve with equal-interval increments was not necessary. For this reason, no attempt was made to equate difficulty levels or content of achievement items. It was arbitrarily decided to cover each 20 percent portion of a course with five achievement items. Moreover, it was reasoned that items covering those parts of a course which had not yet been presented would not be answered correctly. Therefore, only items based on topics previously presented in class were administered at any time. Following this procedure, five achievement items were added to the achievement test following each 20 percent of the course. The final achievement test included twenty-five items, five of which covered the topics presented in the final 20 percent of the class, and twenty of which covered all topics presented previous to the final 20 percent of the class. A similar pattern was followed for each of the five achievement tests: five new items were added to cover the immediately preceding portion of the course, and five other items were chosen to cover other material presented in each preceding 20 percent portion of the course. The first achievement test was based only on standardized scores obtained from instructor-constructed quizzes. The faculty coordinator for each course chose items which were representative of material presented in class for each succeeding achievement test from the established item pool. Thus, the number of items at each successive administration was five (based on standardized score), 10, 15, 20, and 25. In each case, items were representative of topics presented in class to that point in the class. Achievement scores were the number of correct responses.

Content validity of achievement measures was assured by the method of item construction and selection. The corrected split-half reliability coefficients for the last four administrations based on data from the previously described random sample of one hundred students reflected the increasing number of items through successive measurements. These reliabilities ranged from .36 for the second administration to .75 for the fifth administration. A corrected split-half reliability coefficient of .63 was also obtained on the final achievement measure for a random sample of twenty-five students from the Missouri class.

It should be noted that in the construction of the attitude toward psychology scale and the achievement items the Missouri instructors concurred that the items were also objectives of their course of instruction.

Data Collection

A prepared statement introducing the study was read to students by section instructors at the first administration of the attitude scale. Attitude scales were administered by section instructors at approximately the end of each 20 percent of the respective courses. The order of administration of the five forms of the attitude toward psychology scale was the same for all classes. Since items within the scales were randomly arranged, counterbalancing of forms did not seem necessary. To assure that student responses would not be biased by the possibility of instructors examining the completed scales, completed scales were returned to a student monitor who sealed all scales for his section in an envelope.

Achievement measures were obtained either as part of a regular class quiz administered in section meetings or as keyed items on regular one-hour examinations. Inclusion of achievement items in the context of regular class testing assured that students would respond to the best of their ability. In all cases, achievement items were drawn from the established item pool. However, the items at any administration were not necessarily the same between class situations, nor were the items necessarily the same for any class topic over administrations within a given class.

Analysis of the Data

Curves were fitted to the attitudinal variable and achievement variable over time. The curve fitting was accomplished by determining the mathematical function which most adequately accounted for the total variation in the variable through a polynomial regression analysis (5). By fitting curves to attitude toward psychology and achievement, it was possible to compare changes in the variables over time and determine if any longitudinal covariation was evident.

The three Introductory Psychology courses were treated as complete replications of the study to test the generalizability of the results. All statistical analyses were programmed for a CDC 6400 computing system.

RESULTS

In order to conserve space, the complete analysis will not be reported. The regression equations resulting from the data for each variable in each of the learning situations were solved, and the resulting predicted scores converted to a percentage of the total possible score so the curves could be plotted on the same coordinates for comparison (Figures 1, 2, and 3). Although the equations for each variable differed somewhat from situation to situation, it will be noted that the resulting curves are of similar shape across learning situations.

Achievement in all three situations rises in nearly a linear manner, although the curves are represented by as high as a fourth-degree equation. Scores on the Missouri final achievement examination were about 10 percent lower than scores on the two CSU final examinations.

Attitude toward psychology was best described by a horizontal mean line in the CSU Winter and Missouri classes. A slightly rising linear function described the attitude change in the CSU Fall class. It is evident that, in the three classes studied, participation in the learning experience did not change attitude toward subject matter.

There appears to be no relationship over time between the two variables in any situation studied.

TABLE 1

MEAN ACHIEVEMENT SCORES BY ADMINISTRATION OF HIGH AND LOW FINAL ATTITUDE TOWARD PSYCHOLOGY GROUPS COMPARED WITH PREDICTED MEAN ACHIEVEMENT SCORES OF THE TOTAL CLASS

Group	Administration 1	2	3	4	5
High (N=38)	2.66	8.97	12.55	16.00	20.23
Predicted	2.62	8.16	12.44	16.12	19.88
Low (N=38)	2.66	8.08	11.94	15.81	19.26

Achievement and Changing Attitude Toward Psychology

At this point it was decided to compare changes in achievement over time of individuals with specific patterns of change in attitude toward psychology. Using the Winter CSU class as subjects, those individuals in the high and low 25 percent of the distribution

FIGURE 1. CSU FALL PREDICTED CURVES FOR ATTITUDE TOWARD PSYCHOLOGY, AND ACHIEVEMENT SCORES.

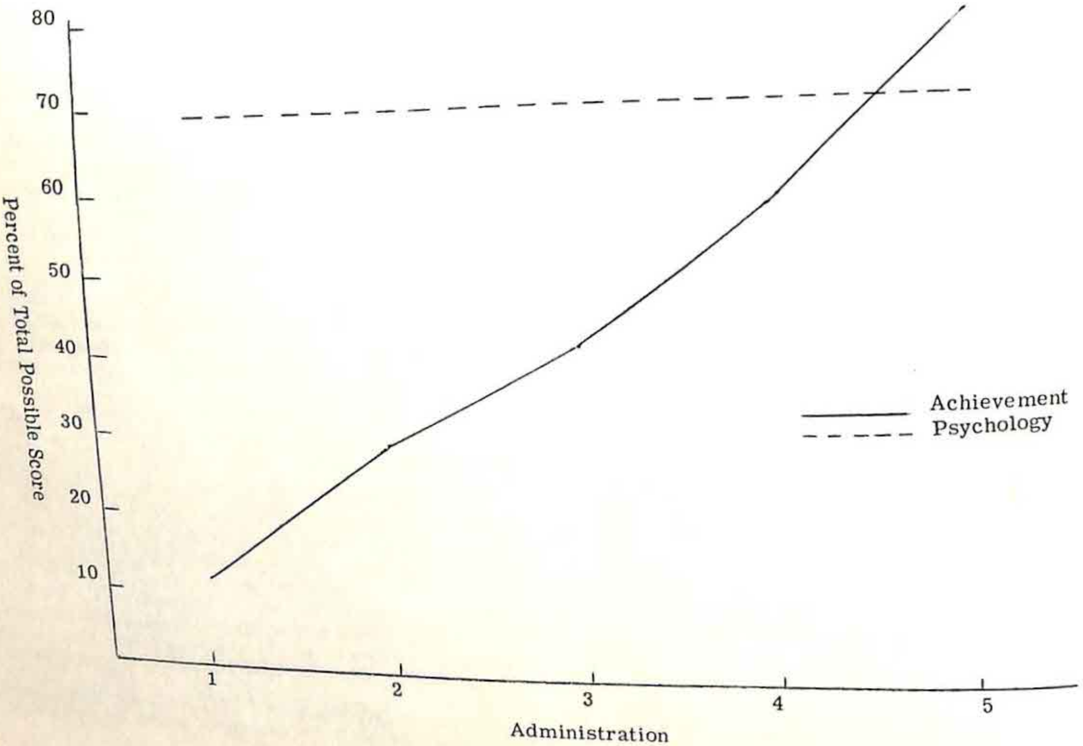
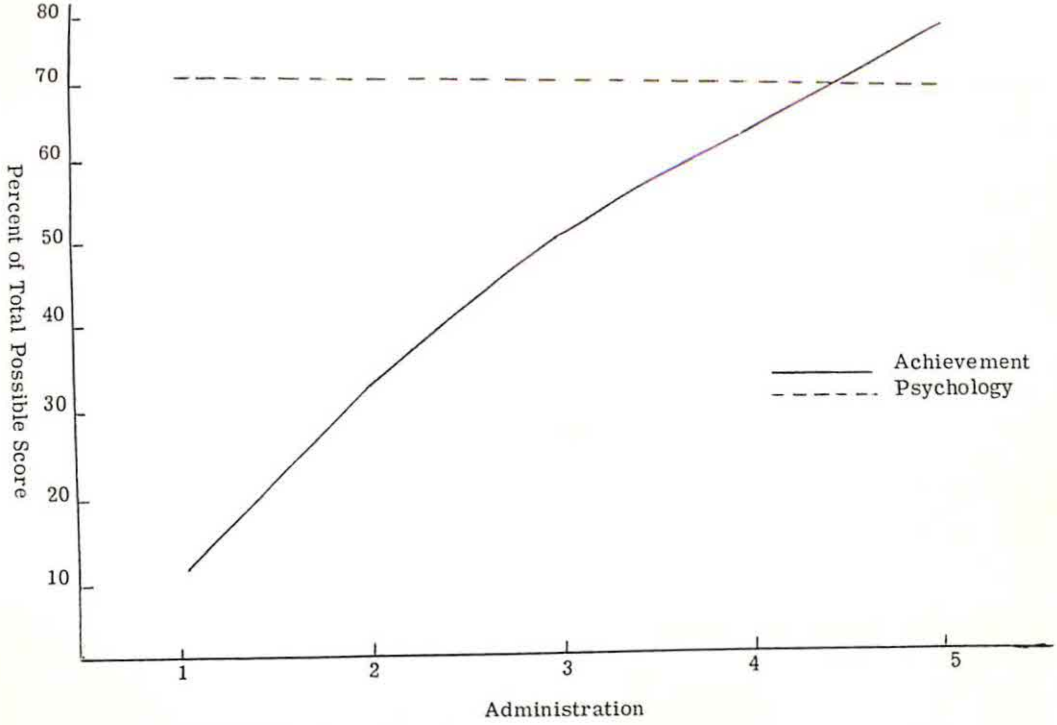


FIGURE 2. CSU WINTER PREDICTED CURVES FOR ATTITUDE TOWARD PSYCHOLOGY, AND ACHIEVEMENT SCORES.



of the fifth attitude toward psychology scores were identified. Within the high and low final attitude groups, the 50 percent of the students nearest the total class mean on the first administration attitude toward psychology scores were identified. A group beginning near the class mean and exhibiting a rising attitude toward psychology (N=38) and a group beginning near the class mean and exhibiting a declining attitude toward psychology (N=38) were thus identified. If affective were related to cognitive learning, the achievement curves for these groups should have differed from each other and from the curve of the total CSU Winter class. Mean achievement scores were calculated for each group and compared to the predicted achievement scores for the entire class. As can be seen from Table 1, the high, low, and predicted achievement means could hardly be closer if completely random groups had been drawn. There is a slight trend for the high attitude group to have higher achievement scores than the low attitude group. However, the differences are not large enough to indicate any real relationship between affective and cognitive learning when the method of choosing the groups is considered.

DISCUSSION AND CONCLUSION

It appears that in the three learning situations studied that the cognitive goals, measured by objective achievement items, were attained. At least there was sufficient evidence from the resulting achievement curves to conclude that performance was considerably better at the end of each class than at the beginning. Affective teaching objectives, however, did not appear to be attained. Attitude toward psychology, measured by attitude items that all instructors agreed they were trying to influence, was, at best, merely maintained throughout a course of instruction.

Contrary to the most common assumptions about the relationship between cognitive and affective learning outcomes, it appears that the two realms are relatively independent. Evidence from this study strongly suggests that changes in one realm do not necessarily lead to changes in the other realm. This conclusion was supported by the similarity of the results in all three learning situations. Further support was given to the conclusion of independence by the finding that students with diverging attitudes toward psychology over time did not differ in cognitive achievement from the predicted achievement of the total class.

The use of only classes taught by closed-circuit television could have affected these results. It may be that affective goals are more difficult to teach through a televised presentation or that the lack of principal instructor-student interaction makes affective involvement with the subject matter harder to achieve. However, the individuals which were identified with diverging attitudes still did not exhibit differential achievement patterns, suggesting that the single method of presentation is not a severe limiting factor.

It does appear that different instructional procedures may have to be employed to achieve both cognitive and affective teaching objectives. The findings of the present study suggest traditional lectures with limited discussions, even though designed to attain certain affective objectives, are not effective in achieving affective change. Further research in the area using an experimental approach with different instructional procedures designed to teach affective objectives is suggested.

NOTE: The research reported herein was one of several projects performed pursuant to a Title VII grant from The Office of Education, U. S. Department of Health, Education, and Welfare.

REFERENCES

1. Aiken, L. J.; Dreger, R. M., "The Effects of Attitudes on Performance in Mathematics," Journal of Educational Psychology, 52: 19-24, 1961.

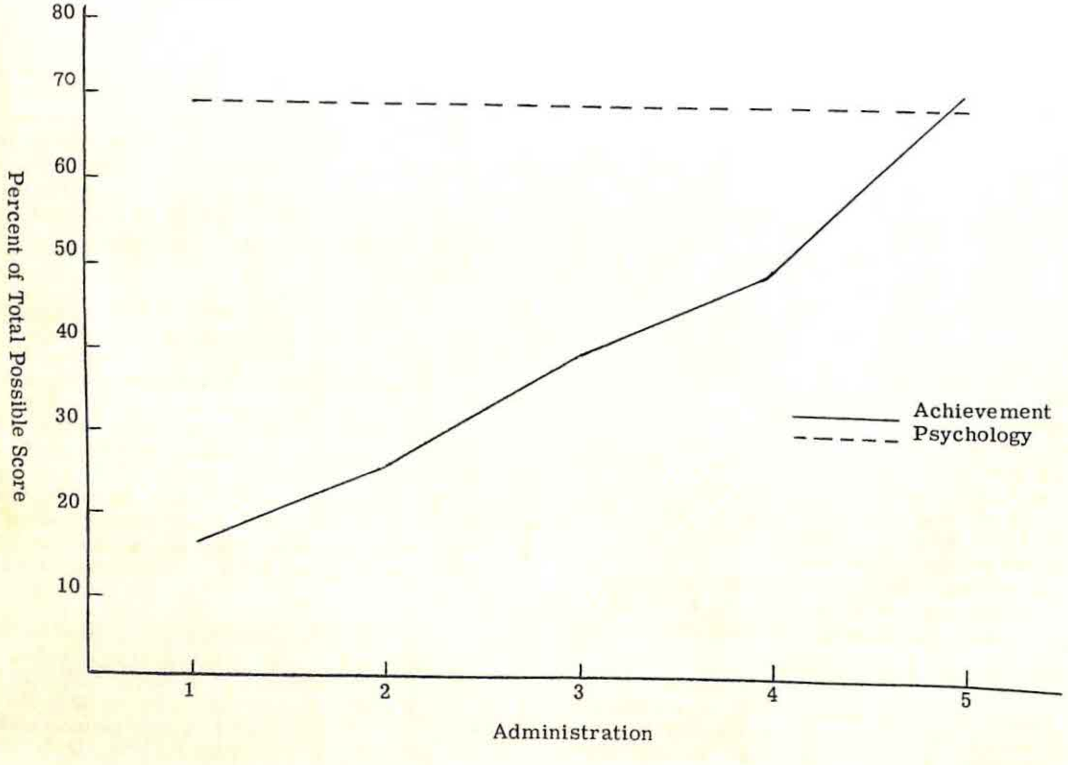
2. Garverick, C. M., "Retention of School Learning as Influenced by Selected Affective Tone Variables," Journal of Educational Psychology, 55: 31-34, 1964.

3. Kratwohl, D. R.; Bloom, B. S.; Masia, B. B.; Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook II: Affective Domain, David McKay Co., New York, 1964.

4. Neidt, C. O.; Hedlund, D. E., "The Relationship Between Changes in Attitudes Toward a Course and Final Achievement," Journal of Educational Research, 61: 56-58, 1967.

5. Ostle, B., Statistics in Research, 2nd. Ed., Iowa State University Press, Ames, Iowa, 1963.

FIGURE 3. MISSOURI PREDICTED CURVES FOR ATTITUDE TOWARD PSYCHOLOGY, AND ACHIEVEMENT SCORES.



THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

LEARNING AND VARYING SEQUENCES OF NINTH-GRADE MATHEMATICS MATERIALS[1]

FRED NIEDERMEYER, JERRY BROWN, and BOB SULZEN

University of California at Los Angeles

ABSTRACT

This experiment compared three sequence versions (logical, scrambled, and reverse frame orders) of the guided discovery program, Number Series, used by Gagné and Brown in their 1961 study. Sixteen ninth-grade algebra students in each of the three sequence groups plus a control group served as the subjects. While the logical order group was the only sequence group to perform significantly better than the controls on both a test of concepts and a problem solving test, none of the three sequence groups differed significantly from each other on posttest performance. Logical group students did, however, make significantly fewer program errors. They also tended to consider the program "interesting," whereas scrambled and reverse order groups felt "neutral" about the program.

THE NOTION of sequence has long been enshrined in the theory of curriculum in general and programmed instruction in particular. Skinner, in his classic article, "The Science of Learning and the Art of Teaching," indicated the logical necessity of sequence: "The device [teaching machine] makes it possible to present carefully designed material in which one problem can depend upon the answer to the preceding and where, therefore, the most efficient progress to an eventually complex repertoire can be made" (11:84). If the frames of a rather sophisticated unit of instruction were presented in a randomly sequenced manner, few people would doubt that the resultant learning would be less efficient and less effective than a carefully planned sequence where "each successive experience builds upon the preceding one . . ." (12:55).

Thus it is a rather perplexing state of affairs when experimenters, using programmed instruction to control sequence and content, are unable empirically to validate the sequence hypothesis. At least six studies over the past 8 years (5, 6, 7, 8, 10, 13) have shown no significant differences in learner achievement when comparing logical and random sequence treatments. In the one study that does support the sequence hypothesis, Roe (9), the results are uncertain. Roe found a significant difference (10 percentage points) favoring a logical sequence when comparing logical and scrambled programs administered on 4 x 6-inch cards in file boxes to college engineering students. Yet another version of the logical program, administered by means of a programmed text format, produced a mean that was six percentage points lower than the scrambled sequence on the cards.

From the remaining studies related to logical and random sequencing of programmed materials (i. e. the "no difference" studies), three hypotheses have evolved:

1. Sequence effects depend upon the subject matter being taught. The more interrelated the tasks are---in the sense of a Gagné type hierarchy (3)---the more effect sequence has on learning.
2. Sequence effects depend upon the individual learner's ability. High ability learners are able to overcome the effects of poor sequencing better than low ability learners.
3. Sequence effects may depend upon the age of the learner. Older, college level learners may be less affected by sequence than younger learners.

The present study examined the effects of logical, scrambled, and reverse order sequences on the learning of a series of mathematical tasks at the ninth grade level. Because of the age of the subjects, and because the materials covered what was thought to be a conceptual hierarachy, it was hypothesized that scrambled and reverse order sequences would be less efficient and less effective (as indicated by instructional time, error rate, and posttest performance) than the logical sequence. It was also hypothesized that posttest differences between high and low IQ subjects would be greater under scrambled and reverse conditions than under logical order.

METHOD

Subjects

Sixty-four ninth-grade students comprising two algebra I classes at a suburban junior high school in the Los Angeles area served as subjects in the investigation.[2] They were randomly assigned to logical, scrambled or reverse order sequence conditions and a control group, making a sub-total of sixteen in each of the four groups. California Test of Mental Maturity (CTMM) Total IQs were available on all but four of the students, and posttest results were available on all but one of the logical students who was absent on the day of the tests. The mean IQ for each group was between 111 and 112, and the variances within each group were not statistically different.

Program

This experiment used a modified version of the Number Series program[3] from Gagné and Brown's 1961 study, "Some Factors in the Programming of Conceptual Learning" (4). The original program consisted of eighty-nine introductory frames followed by forty guided discovery frames which required the learner to "actively produce... concepts... which he had just previously learned in the introductory program" in order to attain the problem solving skill of "finding a formula for the sum of n terms in a number series" (4:319). This particular learning program was selected because it was thought to be based on the knowledge hierarchy validated for this same terminal task by Gagné in his 1961 article, "The Acquisition of Knowledge" (3).

Initial tryout of the program with individual learners found them rapidly "extinguishing" on the eighty-nine introductory frames. Twenty-eight frames were then systematically removed. (For example, if the item, "The term-value for $n = 5$ is ____," appeared six times, with different numerals for n, then two of these practice frames would be removed.) This modified, 61-frame version proved equally as effective as the 89-frame introductory program with three different learners. Nine new frames, five for giving extra practice on one of the concept symbols (T), and four designed to provide a model of the terminal problem solving performance, were inserted between the sixty-one introductory frames and the forty guided discovery frames to make a total of 110 frames.

Individual frames were dittoed on 5 x 9-inch cards with answers on the backs of the cards. Frames were scrambled by use of a table of random numbers, and a different scrambled sequence was developed for each of the sixteen students in that group. Reverse order sequences were formed by merely arranging the cards of a logical sequence from back to front.

Procedure

Approximately twenty-three experimental students in each of the two algebra classes worked at their desks in the presence of the three experimenters about 45 minutes a day for two consecutive days. Each student had the 110 cards stacked in a covered folder on his desk. He was allowed to work on only one card at a time and inserted each finished card into a second folder. Students were given red ink pens to write with and were told to draw a slash mark through any blank they could not answer before turning the card over to see the correct answer. By focusing the results of their activity on helping the experimenters obtain data to revise the program (all connotations regarding student grades were removed), students readily "played the game," and no student was observed to be "cheating" during the two days of instruction.

A posttest only, control group design was employed. Control students were assigned to the library during instruction, but returned on the third day for testing.

Measures

Independent variables were the three sequence treatments and IQ. The dependent variables consisted of *time to complete the program*, *errors made in the program*, *scores on a test of the concepts* taught in the introductory part of the program, *scores on a transfer test* of the terminal problem solving task (finding formulae), and an *attitude measure*. The test on the concepts presented in the sixty-six introductory frames consisted of ten tasks taken directly from the program (For example, "If $T_n = 9$, $\sum^{n} =$ ______"), but applied to a new number series.

The reliability coefficient for this test, calculated by Stanley's Split-Half Procedure (1:316), was .82.

The transfer test of problem solving consisted of five new number series for which the student was expected to state a general equation that would give the sum of the series at any term, n. Three of these series were used as test items by Gagné and Brown, and the other two were developed by the experimenters. Responses to each of the five series were scored on a two point basis for a total of ten points. This test was found to be exceedingly difficult for the students, and no attempt was made to assess its reliability in the present study.

RESULTS AND DISCUSSION

Program Measures

Table 1 presents means and standard deviations for time to complete the program and errors on the program for each sequence condition. Analysis of variance indicated that program error rates differed significantly (.01) in favor of logical sequencing, but instructional time differences did not. The significant error rate differences are taken to indicate the existence of a high degree of sequencial dependence among the frames in the logical version. Apparently correct responses to some of the items depended upon

TABLE 1

TIME TO COMPLETE PROGRAM AND ERRORS ON PROGRAM FOR EACH SEQUENCE CONDITION (N = 16 FOR EACH GROUP)

Sequence Condition	Time (Minutes) M	Time (Minutes) SD	Program Errors M	Program Errors SD	Program Errors Error Rate
Logical	78.5	15.3	38.2	16.6	35%
Scrambled	80.9	13.6	54.8	17.0	50%
Reverse	81.5	6.5	52.1	13.8	47%

knowledge of previous items along the lines of a knowledge hierarchy delineated by Gagné (3). Randomizing or reversing this logical sequence caused students to make significantly more program errors. This probably would not have happened if the items had been independent of each other. A list of discrete facts about the anatomy of the ear or a list of spelling words should generate the same error rate no matter how the items are sequenced.

An examination of program errors broken down into blocks of ten frames for each of the sequence conditions reveals some interesting patterns. Figure 1 contrasts program errors for the students under the logical condition. The erratic rise and fall of the logical line can be explained in terms of the program content. The first fifty frames present to the learner the introductory concepts (term number, term-value, sum) needed for the problem solving skill taught in the last forty frames. The mean number of errors in each block of ten frames rises steadily as students begin to "drop out" in the conceptual hierarchy. Errors decrease for the next twenty frames (51-70) because these frames review introductory concepts and provide a model of the terminal problem-solving performance. The last four blocks of ten frames (71-110) correspond to Gagné's guided discovery program in which he leads students to "discover" the formula to each of four different number series. The sharp decrease in errors on the second series (frames 81-90) is probably due to the fact that this particular series has a relatively simple formula (n^2). The means for the scrambled sequences, as would be expected, show very little variance and tend to decrease slowly as more and more information becomes available to students.

Figure 2 contrasts program errors for the logical sequence with those of the reverse sequence

FIGURE 1

MEAN NUMBER OF ERRORS PER BLOCK OF TEN INSTRUCTIONAL FRAMES FOR THE LOGICAL AND THE SCRAMBLED SEQUENCE CONDITIONS (N = 16 FOR EACH GROUP)

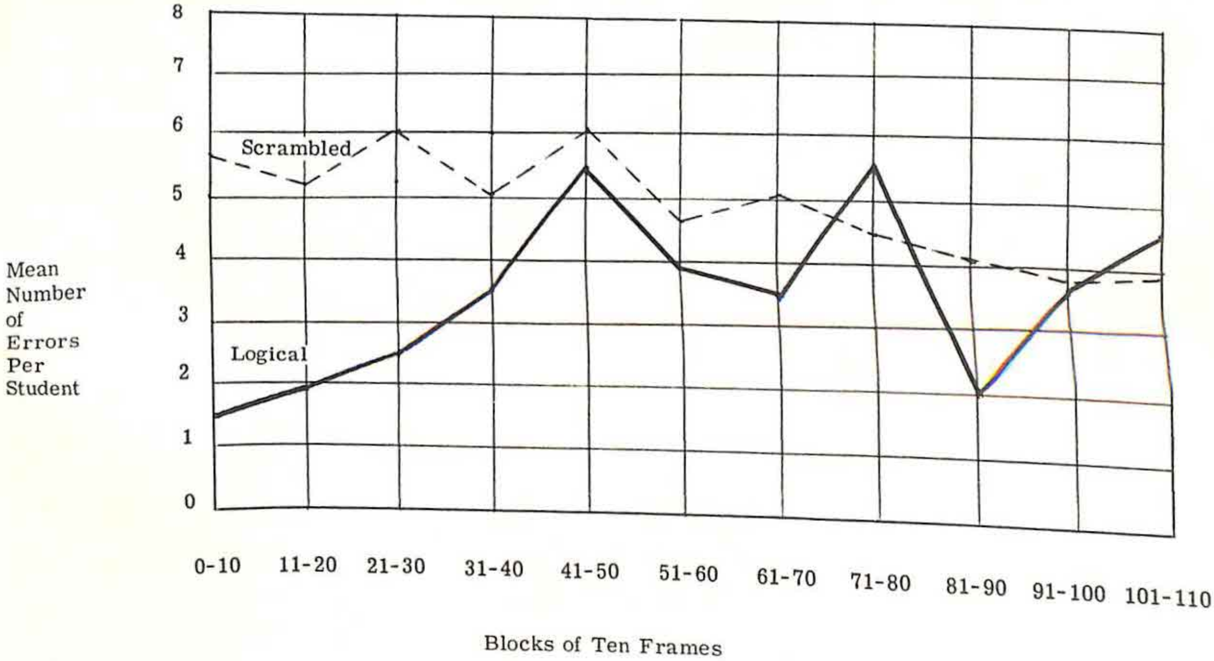
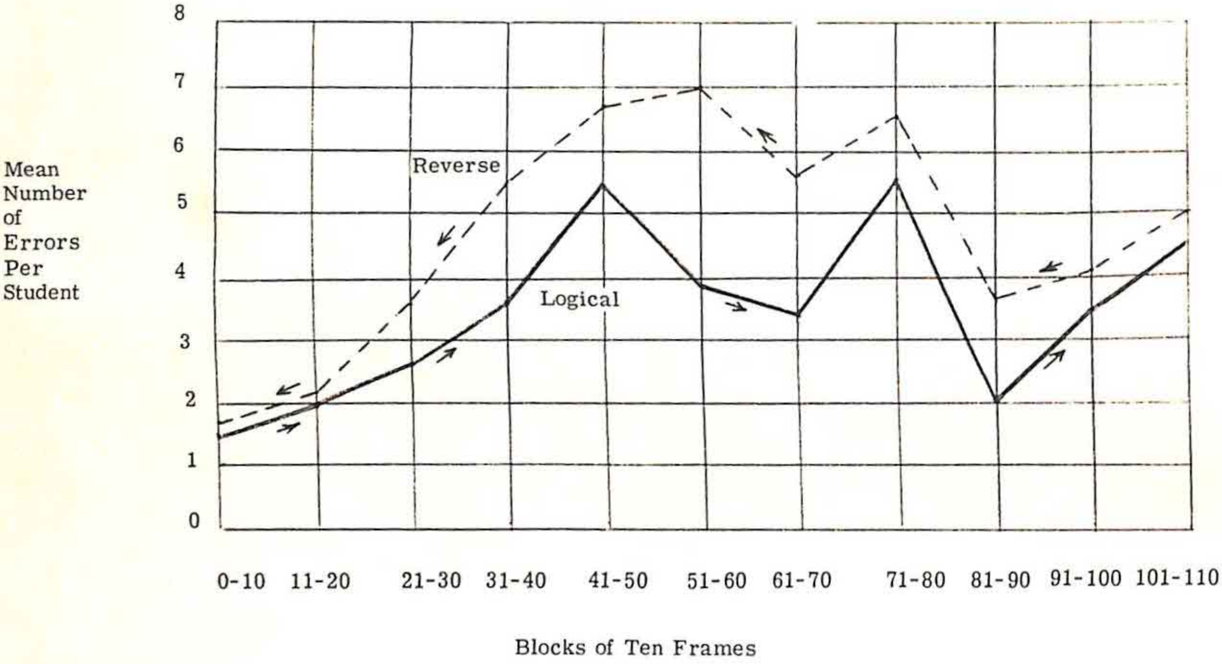


FIGURE 2

MEAN NUMBER OF ERRORS PER BLOCK OF TEN INSTRUCTIONAL FRAMES FOR THE LOGICAL AND THE REVERSE SEQUENCE CONDITIONS (REVERSE SEQUENCE HAS BEEN TURNED AROUND TO CORRESPOND TO LOGICAL SEQUENCE. ARROWS INDICATE DIRECTION IN WHICH STUDENTS WORKED THROUGH PROGRAM. N = 16 FOR EACH GROUP)



turned around. In other words, errors of the reverse order students are shown from frames 1 through 110 although the students worked these frames beginning with frame 110 and ending with frame 1. It will be noted how well the reverse order error rate, when turned around to correspond to the logical order, matches the logical order "dip-for-dip," but at a consistently higher level of errors. The striking similarity between the two lines was rather surprising to the authors. One possible explanation is, of course, that many of the frames in the program are over-cued.

Post Program Measures

Means and standard deviations of performances relative to the two posttests for each sequence condition are given in Table 2. For both tests, analysis of variance indicated significant differences among the means ($p < .01$). Posteriori comparisons, however, using the Scheffé method (2:296) to hold the probability of a Type I error constant, reveal that only the logical sequence group performed significantly better than the control group on both tests. The scrambled sequence group performed significantly better than controls on the Test of Introductory Concepts but not on the Transfer Test. The reverse sequence group performed significantly better than controls on the Transfer Test, but not on the Test of Introductory Concepts. None of the three sequence groups differed significantly on either test

TABLE 2

PERFORMANCE MEASURES ON TEST OF INTRODUCTORY CONCEPTS AND TRANSFER TEST OF PROBLEM SOLVING SKILL FOR EACH SEQUENCE CONDITION AND CONTROL GROUP

		Measure			
		10 Item Test of Introductory Concepts		10 Point Transfer Test of Problem Solving Skill	
Sequence Condition	N	M	SD	M	SD
Logical	15	5.5	3.1	3.3	3.0
Scrambled	16	4.1	2.9	1.8	3.0
Reverse	16	3.4	3.5	2.6	2.1
Control	16	0.8	1.0	0.3	0.7

from each other. The Kruskal-Wallis rank test also failed to reveal any statistical differences between the logical, random, and reverse sequence groups (2:362).

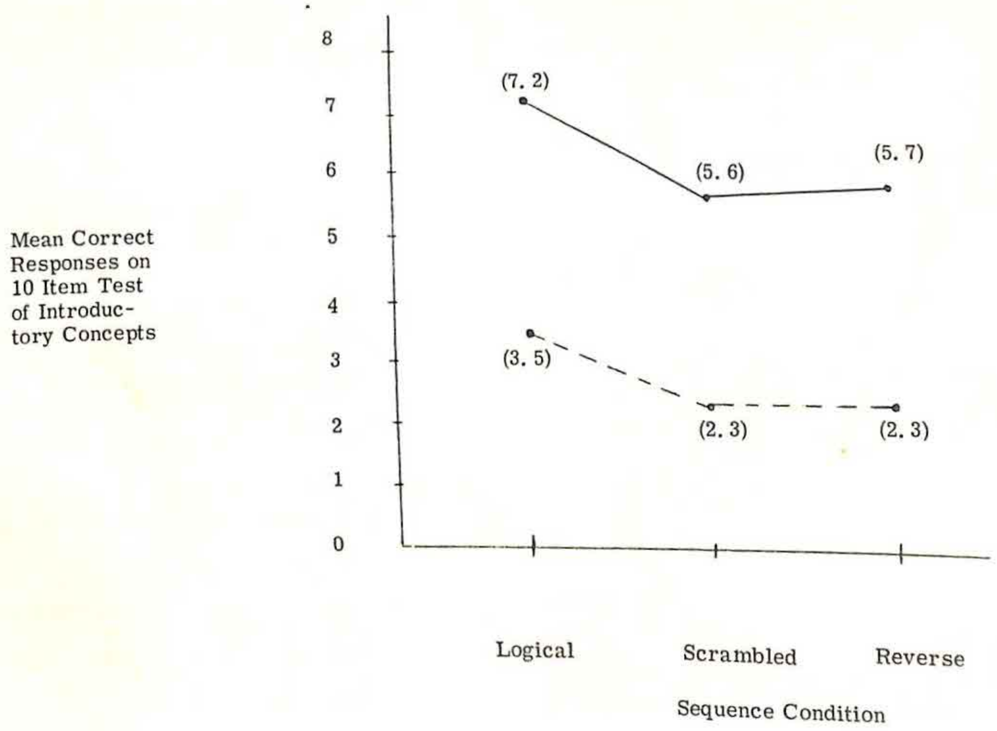
The possibility of IQ and sequence interaction effects on the Test of Introductory Concepts was examined by means of a factorial analysis of variance. Non-significant effects for sequence and sequence x IQ interaction were indicated. The F ratio for IQ, however, was 4.68 ($p < .05$). Figure 3 presents means of each sequence group on the Test of Introductory Concepts after the groups had been split at the median IQ. The parallel lines indicate little or no interaction effect, and the higher means favoring the logical sequence group are not significant. It seems that aptitude, rather than sequence, determined how well an individual learner would do under all three sequence conditions.

An attitude measure administered by the classroom teacher the day following the study revealed an interesting difference. Students in the logical sequence group tended to consider the program "interesting," whereas students in the scrambled and reverse treatment groups tended to feel "neutral" about the program (Chi-square = 6.6, $p < .05$). No differences were found relative to how difficult the students felt the materials and the tests were. That the logical students felt the program was more interesting than scrambled and reverse order students would be expected because their error rate on the program was significantly less. The program was not "punishing" for students under the logical sequence condition. It is also possible the logical ordered students professed higher affect toward the program because they were able to follow the simple-to-complex development from concepts and principles to problem solving skills, whereas students in the other two groups never were able to figure out what the program "was driving at" even though they learned to perform some of the tasks taught in the program.

The data of this study reveal no significant differences between logical, random, and reverse order sequence conditions relative to instructional time and posttest performances (although it should be remembered that the logical group students were the only ones who performed significantly better than non-instructed controls on both posttests). The expected interaction between sequence and IQ also failed to materialize. In this study there were no significant differences between sequence groups relative to cognitive outcomes. The significant difference in learner affect, however, may be enough to

FIGURE 3

MEAN NUMBER OF CORRECT RESPONSES ON TEST OF INTRODUCTORY CONCEPTS BY HIGH AND LOW IQ STUDENTS UNDER LOGICAL, SCRAMBLED, AND REVERSE SEQUENCE CONDITIONS (N's EQUALED: HiIQ-LOGICAL = 7, HiIQ-SCRAMBLED = 9, HiIQ-REVERSE = 6, LoIQ-LOGICAL = 6, LoIQ-SCRAMBLED = 7, LoIQ-REVERSE = 9)



justify careful attention to sequence in short instructional programs.

In reflecting upon the results of the investigation, the authors have come to the same conclusion Gagné has advocated in many of his studies: What is learned is more crucial in learning effectiveness than how it is learned (4:320). It must be remembered that all subjects responded to the same stimuli (frames) even though the frames were presented in various orders. The students in the scrambled and reverse order sequence groups apparently were capable of integrating and organizing information regardless of the sequence of presentation. Instead of studying the effect of frame sequence in short instructional programs, perhaps more could be learned by studying the effects of deleting key frames related to certain tasks in a validated knowledge hierarchy and replacing these frames with placebo frames. (This was one of the original intentions of the present study, but, unfortunately, the Number Series program appears to have been written before the knowledge hierarchy in "The Acquisition of Knowledge" was derived by Gagné, and the present investigators were unable to find much correspondence between the program and the hierarchy.) Though sequence effects would be expected to appear in fairly long instructional programs, the present study, as well as previous studies, lead the authors to believe that for short (less than two hours) programs, sequence may not be as crucial to cognitive outcomes as has previously been thought.

FOOTNOTES

1. Paper presented at the California Educational Research Association Annual Meeting, Berkeley, California, March, 1968.

2. The cooperation of the Montebello Unified School District during the study is greatly appreciated by the authors.

3. The authors are indebted to Dr. Gagné for allowing them to use the learning program, Number Series.

REFERENCES

1. Ebel, Robert L., Measuring Educational Achievement, Prentice-Hall, Inc., New Jersey, 1965.

2. Ferguson, George A., Statistical Analysis in Psychology and Education, McGraw-Hill, New York, 1966.

3. Gagné, Robert M., "The Aquisition of Knowledge," Psychological Review, 69:355-365, 1962.

4. Gagné, Robert M., and Brown, L. T., "Some Factors in the Programing of Conceptual Learning," Journal of Experimental Psychology, 62: 313-321, 1961.

5. Hamilton, Nancy R., "Effects of Logical Versus Random Sequencing of Items in an Autoinstructional Program Under Two Conditions of Overt Response," Journal of Educational Psychology, 55:258-266, 1964.

6. Levin, G. R., and Baker, B. L., "Item Scrambling in a Self-Instructional Program," Journal of Educational Psychology, 54:138-143, 1963.

7. Miller, Herbert R., "An Investigation into Sequencing and Prior Information Variables in a Programed Instruction Unit for Junior High School Mathematics," (Mimeograph), University of Southern California.

8. Payne, D. A., Krathwohl, D. R., and Gordon, J., "The Effect of Sequence on Programed Instruction," American Educational Research Journal, 4: 125-132, 1967.

9. Roe, A., "A Comparison of Branching Methods for Programed Learning," Journal of Educational Research, 55:407-416, 1962.

10. Roe, K., Case, H., and Roe, A., "Scrambled Versus Ordered Sequence in Autoinstructional Programs," Journal of Educational Psychology, 53: 101-104, 1962.

11. Skinner, B. F., "The Science of Learning and the Art of Teaching," in Teaching Machines and Programed Learning, Lumsdaine and Glaser, (eds.), National Education Association, 1960, pp. 75-99.

12. Tyler, Ralph, Basic Principles of Curriculum and Instruction, University of Chicago, 1950.

13. Wodtke, K., Brown, B., Sands, H., and Fredericks, P., "Scrambled Versus Ordered Sequencing in Computer-Assisted Instruction," USOE Project No. 1435, Pennsylvania State University, July, 1967.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

ASSESSMENT OF MATHEMATICAL CONCEPTS OF FIVE-YEAR-OLD CHILDREN

ANTHONY N. SCHWARTZ

State University of New York, Plattsburgh

ABSTRACT

An instrument, the Schwartz Early Mathematics Inventory (SEMI), was devised to assess, by means of group administration, the mathematical achievement of children ages 3 through 5. The data presented concern only the first stages of the process and deal only with 215 kindergartners. A reliability of .94 was obtained on the eighty-one items administered on 64 pages. The mean for items correct 43.2. The proportion of correct responses for various items ranged from .85 to .12. The children had satisfactorily acquired the physical and emotional attributes necessary for completing the inventory, namely, holding a pencil or crayon, turning pages, recording responses in designated areas, and maintaining satisfactory attention for the 12 to 28 minutes required for administering each of the three portions of the SEMI.

THE DEVELOPMENT of procedures for the assessment of the capabilities and the measurement of accomplishments of five-year-olds is a major challenge to the educational world. The Research and Development Center in Educational Stimulation, University of Georgia, has unified its total program of activities around early and continuous intellectual stimulation of children, ages 3 through 12. To carry on such a program requires a sound basis of evaluation for making decisions.

The principal purpose of this research was to develop an instrument which could systematically assess the variations in performance, involving the acquired mathematical concepts and skills of kindergarten children.

BACKGROUND FOR THE STUDY

Neither child growth and development specialists, mathematicians, nor professional educators seem to be in accord as to what mathematical concepts a child has developed at the preschool level or what the curriculum should encompass at this age. Research has been limited in scope and the results do not lead one to make conclusive decisions. Opinions differ, some lead one to believe that the child has a very superficial concept of numerosity while others, pertaining to children of a socially biased culture, state that an extensive background of mathematical acumen has been acquired prior to that developed by the organized school program.

The studies of the developmental stages by Piaget (9) serve as a basis for much discussion and argument concerning the mathematical abilities of children, including ages 4 and beyond. Dodwell (2 and 3), Holmes (6), Inhelder (7), Elkind (4), and Wohlwill (12), have done replications of Piaget's studies and have tended to support his findings, but in so doing these researchers have had to improve upon the investigative procedures. Research by Estes (5), however, refutes many of the findings mentioned. The author suggests a careful scrutiny of the book by Wallace (10) as a source of not only the studies mentioned but many others which are relevant. Although most of these studies indicate some evidences of stages of development, no agreement appears.

Recent studies substantiate the statements that children come to kindergarten with considerable knowledge upon which the school can build an interesting, challenging, and sequential curriculum. Studies by Sister Josephina (8) and Brace and Nelson (1) involved individual evaluations of children. The first studied thirty children in one kindergarten, and the second used a sampling of 124 children drawn from a population of 3,000. Williams (11) administered a group test to 595 kindergarten entrants and reported a positive response ranging from 6.9 to 81.3 percent of the sixty-three items given.

As far as could be ascertained there are no commercial tests available to specifically measure the cognitive mathematical status of children ages 3 to 6 or to evaluate progress as they are intensively subjected to formal instructional procedures.

PURPOSES

The main purpose of this investigation was to develop a group pencil and paper instrument to assess the mathematical achievement of kindergarten children. Secondary concerns were: (a) assessment of the physical requirements involved in the procedures, such as turning pages, holding a pencil, and attention span, (b) assurance that the response of the individual can be accurately recorded, and (c) evidence that the response is individual rather than a "cooperative" endeavor.

DEVELOPMENT OF THE INSTRUMENT

In preparation for the development of the Schwartz Early Mathematics Inventory (SEMI)[2] three primary sources were used to develop concepts and the initial item pool. First, a literature survey was made of material regarding individual or group testing which might have implications for measuring mathematical achievement. Tests, whether general or for specific subject areas, were investigated regarding such areas as: format, manner of presentation, giving directions, and recording pupil response. Available materials and workbooks, methodology books, early childhood curricula, and general child growth and development releases were also included. A panel of educators reviewed the items.

Second, the author conducted individual assessments of 268 kindergartners.

Third, kindergarten teachers were interviewed in respect to their number experiences with children.

Since small muscle control must be taken into consideration, it was determined to use but one vertical line, in a majority of situations, for responding to each item. Whenever feasible sets were represented by circles and squares so that required vocabulary would be reduced for those with backgrounds of cultural deprivation.

As a result, eighty-one items[4] were selected having a range from easy to difficult within specific areas. The areas ranged from simple to complex in mathematical learnings which might have probable relevance to this age group. Items were presented in numerous ways to give variety and to ascertain adaptability of the children.

Pre-inventory exercises of approximately 15 minutes duration were given to the children in order to familiarize them with the social setting, the test carrels and mechanical preocedures. The children were asked to mark with pencil or crayon, whichever was most frequently used in the classroom, and were instructed to void any marks they did not want by making two horizontal marks across the vertical mark they had made. The double line was used to differentiate from X which many children were prone to use. In general, the test booklets contained one item per page--eighty-one items on 64 pages--to avoid complicating the task of responding by requiring even rudimentary reading skills.

SUBJECTS

Complete data were obtained from 215 kindergartners residing in northern Georgia whose age range was 5 years 0 months to 6 years 5 months. The children attended public and private schools from varied socio-economic, racial, and ethnic backgrounds as follows:

a. rural, low socio-economic children were obtained from the total kindergarten enrollment of one county in three schools, one of which was Negro. Total N = 66.

b. suburban children were obtained by the stratified sampling of the total enrollment of 5-year-olds in a county in a metropolitan area of approximately one million population, proportion 9 percent Negroes. Total N = 57.

c. urban, low economic children were obtained from the total enrollment of three schools, selected under a Title I grant as being culturally deprived, in a city of 18,000. Approximately 50 percent of this group were Negro and were enrolled in one school. Total N = 60.

d. urban, high socio-economic culturally advantaged children were obtained from the total enrollment in a private, all white kindergarten. Total N = 32.

ADMINISTRATION OF THE SEMI

The children were examined during the first 3 weeks of February in rooms and hallways with which they were familiar, and in their own classrooms if their peers were assigned to carry on activities in another place. The children were seated six at a table, separated by "carrels" of interlocking one-fourth inch plywood 1-foot high, with a minimum work space of 15'' x 32''. A maximum of twelve children were tested at any one period. Each group of six children were supervised by either the administrator, a graduate student or a teacher aide. The tests were administered by graduate students or the author but not by the classroom teacher, although she may have been an observer at an administration. The only assistance given during the time the inventory was being administered was to see that the child was on the correct page or to see that he understood where the response was to be placed.

During the approximately 15 minute pre-inventory exercises, the administrator or the assistant gave any aid required or requested which would help the child in understanding the activities and the task to be accomplished.

The items for the SEMI were read once to the children unless the task was somewhat involved or the statement lengthy. The tempo of administering the items was as rapid as feasible to eliminate "peeking," talking and other distracting incidents. The average of two items per minute was suggested as a guide. In order to reduce fatigue, the SEMI was administered on separate days, in three sittings besides the pre-inventory session.

PRESENTATION OF THE DATA

The SEMI booklets were hand scored on a right or wrong basis regardless of the number of foils for the item. No partial credit was allowed. Responses were placed on data cards and computer processed by the TSSA Program (13).

INVENTORY STATISTICS

The general test data for the SEMI is presented in Table 1 as computed from data gathered from 215 kindergartners. The data indicates that the populations did not fall into the normal curve pattern but are leptokurtic and skewed--a rather common outcome with some kinds of empirical data. The kurtosis is significant at the .05 level. The inventory reliability equals .94 using the Kuder-Richardson Formula 20.

The mean inventory score was 43.19. The scores earned by the children ranged from a high of 70 to a low of 8. The latter child omitted fifty-four items. Fifty-two children did not omit any items. The range of administration time for the SEMI was: part I, 12-25 minutes; part II, 10-18 minutes; and part III, 12-28 minutes.

ITEM ANALYSIS

An analysis of the eighty-one items of the SEMI is summarized in Table 2. The difficulty level of each item is indicated in terms of the percentage of correct responses. The proportion of the children selecting a correct response for particular items ranged from .85 to .12. Seven items can be considered comparatively easy as they were answered correctly at least by 80 percent of the children. Eleven items can be considered relatively hard as they were answered correctly by less than 30 percent of the children. No items fell within the extremely easy ($p > .95$) and the extremely difficult ($p < .05$) range.

TABLE 1

MOMENTS OF THE SEMI

Mean:	43.19 ± 1.02
Standard deviation:	14.97
Skewness:	-0.42
Kurtosis:	-0.53 *

* Significant at .05 level

Point-biserial correlations were computed for each item and total test score. They ranged from .62 to .11. Items with point-biserial coefficients above .40 were considered to discriminate effectively. The coefficients for forty-nine items fell into this category.

FACTOR ANALYSIS

Although the SEMI developed as a single-score general mathematics inventory, inter-item tetrachoric correlations were processed by factor analysis with varimax rotation. These inter-item correlations varied from .82 to random negative values as large as -.16. The number of factors to be extracted was programmed at ten. The resultant rotated factor loadings ranged from .80 to -.41 but 84 percent of the items had their highest loadings on the unrotated first factor and 52.2 percent of the common variance is accounted for by this factor, so the data did not lend themselves to interpretation by multiple discrete factors. Factor loadings of .40 or greater, on the rotated factors were given greatest emphasis in assigning the items to test parts and sub-parts. By combining these data with a rational analysis of the inventory, the items were categorized into fourteen comparatively recognizable and workable parts or sub-parts as given in Table 2. (See page 74).

GENERAL OBSERVATIONS

The $8\frac{1}{2}$'' x 7'' format was adequate for the illustrations and materials presented. Presentation of the items in mimeographed two dimensional form did not appear to be a barrier to understanding what was expected.

TABLE 3

ITEMS CONSIDERED COMPARATIVELY HARD--PERCENT OF CHILDREN ANSWERING CORRECTLY $\leq$ 30

Item	Percent
24. Frame has more squares than circles, mark extra squares	24
31. Mark the frame that shows 1/3 of a pie	24
33. Mark the jar that is $\frac{1}{4}$ full of water	12
37. Mark each block/thing which has some black and some white	28
38. Set of different shapes, mark those with more than two sides	14
45. Seriation--sets of lines, make marks (3) for missing set	26
46. Seriation--sets of lines, make marks (4) for missing set	21
54. Bar graph, fill in or mark to look right	14
58. Make 2 more marks than circles (2 circles)	20
80. When I count 5, 4, 3, 2, mark number I should say next	29
81. When I count 10, 9, 8, 7, mark number I should say next	29

The response to most items was usually immediate and the answer readily designated. Instructions were easily remembered and very little repetition was necessary. The cross-out technique for voiding responses proved invaluable and highly suitable for this age group.

SUMMARY AND GENERALIZATIONS

An instrument, the Schwartz Early Mathematics Inventory (SEMI), was devised to assess, by means of group administration, the mathematical achievement of children ages 3 through 5. The data presented here concern only the first stages of the process and deal only with 215 five-year-olds.

A reliability of .94 was obtained on the eighty-one items administered. The mean for items correct was 43.2. The proportion of correct responses for various items ranged from .85 to .12.

The children satisfactorily acquired the physical and emotional attributes necessary for completing the inventory, namely, holding a pencil or crayon, turning pages, recording responses in designated areas, and maintaining satisfactory attention for the 12 to 28 minutes required for administering each of the three portions of the SEMI. By using portable carrels and the usual kindergarten tables, a very high degree of individual response was maintained.

The data are encouraging in light of the purposes of the investigation. Further refining and decreasing the number of items should yield a more usable instrument. Since the youngest children taking the SEMI were able to perform the physical functions necessary, the author is encouraged to validate the instrument for three-and four-year-olds.

FOOTNOTES

1. The research reported herein was performed as a special project as a portion of the postdoctoral education training program sponsored as part of the activities of the Research and Development Center in Educational Stimulation at the University of Georgia pursuant to a contract with the U.S. Department of Health, Education and Welfare grant No. OREG 2-6-061881-1406.

2. A complete copy of the Schwartz Early Mathematics Inventory and related materials are on file at the Research and Development Center in Educational Stimulation, University of Georgia.

3. Grateful acknowledgment is made to Dr. Warren G. Findley and staff of the Research and Development Center in Educational Stimulation, for the assistance offered in conjunction with the postdoctoral training program.

4. Numeral number connotations are used interchangeably because of usage by children, parents, and some teachers.

REFERENCES

1. Brace, B. and Nelson, L. D., "The Preschool Child's Concept of Number," The Arithmetic Teacher, 12: 126-33, February 1965.

2. Dodwell, P. C., "Children's Understanding of Number and Related Concepts," Canadian Journal of Psychology, 14: 192-205, 1960.

3. Dodwell, P. C., "Children's Understanding of Number Concepts: Characteristics of an Individual and Group Test," Canadian Journal of Psychology, 15: 29-36, 1961.

4. Elkind, D., "The Development of Quantitative Thinking: A Systematic Replication of Piaget's Studies," Journal of Genetic Psychology, 98: 37-46, 1961.

5. Estes, B. W., "Some Mathematical and Logical Concepts in Children," Journal of Genetic Psychology, 88: 219-222, 1956.

6. Holmes, E. E., "What Do Pre-First Grade Children Know About Numbers ?" Elementary School Journal, 63: 397-403, April 1963.

7. Inhelder, B. in Wallace, J. G. Concept Growth and the Education of the Child, King, Thorne and Stace Ltd., School Road, Hove, Sussex, British Isles, p. 82, 1965.

8. Josephina C. S. J., Sister, "Quantitative Thinking of Preschool Children," The Arithmetic Teacher, 12: 54-55, January 1965.

9. Piaget, J., The Child's Concept of Number, Routledge and Kegan Paul, London, 1952.

10. Wallace, J. G., Concept Growth and the Education of the Child, King, Thorne and Stace Ltd., School Road, Hove, Sussex, British Isles, 1965.

11. Williams, A. H., "Mathematical Concepts, Skills, and Abilities of Kindergarten Entrants," The Arithmetic Teacher, 12: 261-268, April 1965.

12. Wohlwill, J. F., "A Study of the Development of the Number Concept by Scalogram Analysis," Journal of Genetic Psychology, 97:345-377, 1960.

13. Wolf, Richard and Klopfer, Leopold, "Test Scorer and Statistical Analysis 2 (with Factor Verimax Rotation)," unpublished report issued by the Computer Program Library, Social Science Division, University of Chicago, p. 37.

TABLE 2

GROUPING AND ANALYSIS OF ITEMS OF THE SEMI

Test Item	Description of item--reflects instructions read to children	Percent * correct	Point ** Biserial Correlation
	PART 1. PROBLEM SOLVING		
A.	Number stories (If-then situations) Example: While demonstrating, "If I have two books and I take one away, then how many will I have left ?"		
63.	2 - 1 = n Objects pictured. Mark the right frame	81	51
64.	5 - 2 = n Objects pictured. Mark the frame	70	49
65.	1 + 1 = n No objects. Make right number of marks	60	37
61.	2 + 1 = n Objects pictured. Mark the frame	60	41
62.	3 + 3 = n Objects pictured. Mark the frame	58	44
67.	2 - 1 = n No objects. Make right number of marks	55	35
68.	5 - 1 = n No objects. Make right number of marks	47	54
66.	2 + 2 = n No objects. Make right number of marks	47	51
B.	Reproduction		
43.	Make 2 marks	79	45
44.	Make 5 marks	68	45
57.	Make the same number of marks as circles (2)	61	51
59.	Make one less mark than circles (2)	52	21
51.	Make marks for combining a frame with a set of 1 circle with a frame with a set of 2 circles	42	45
58.	Make 2 more marks than circles (2 circles) [Identical items on different pages]	20	37
60.	Make 2 more marks than circles (2 circles) [Identical items on different pages]	20	33
	PART II. VISUAL DISCRIMINATION		
A.	Matching and Analyzing Shapes		
13.	Mark two numbers or figures that are the same	82	53
14.	Mark two numbers/figures that are the same	78	62
28.	Mark something that measures table/how tall you are	66	46
27.	Mark something that measures days of month	62	42
55.	Mark the two shapes that look alike	74	54
56.	Mark two figures that look alike turned the same	39	42
B.	Recognition of sub-sets		
9.	Mark frame that has 3 squares	81	50
41.	Mark 3 of the squares in the frame	77	48

Description of item--reflects instructions read to children	Percent * correct	Point ** Biserial Correlation
Two frames, if one more than other mark it. If same number of circles in both frames mark each frame	75	48
Mark frame that has 5 squares	75	47
Mark 4 of the circles in the frame	73	58
Mark $\frac{1}{2}$ set of 4 squares (scored two ways)	51	37
Analyzing patterns or classification		
Frame of sticks or heavy lines, mark the stick that doesn't fit	78	39
Frame of circles, mark the circle that doesn't fit	74	44
Four frames, mark the two frames that have the same number (3 squares)	59	65
Four frames, mark each of the two frames that look the same (four pictures of ducks)	54	60
Four frames, mark the two frames that have the same number (9 circles)	53	61
Frame alternate circles-squares as beads, mark circle or square which will make look right	51	31
Frame, square, two circles, square, etc. as beads, mark circle or square which added looks right	50	45
PART III. COMPARISON		
Semi-concrete non-equivalent sets		
Mark the largest dog	85	37
Mark the shortest pencil	84	40
Which frame has more buttons ?	82	44
Draw a line from dot to dot (dots 7 inches apart)	80	44
Which frame has the most pictures ?	77	45
Abstract non-equivalent sets		
Mark frame with set greater than 3	73	38
Frame has more squares than circles, mark extra squares	55	29
Mark frame with set equal to 3	41	24
Which frame has the least number of circles ?	36	40
Mark frame with set less than 3	36	39
Frame has more squares than circles, mark extra squares	24	26

Test Item	Description of item--reflects instructions read to children	Percent * correct	Point ** Biserial Correlation
C.	Ordinality		
17.	Point to left. Mark the first duck.	84	45
18.	Mark the last duck.	67	58
19.	Mark the third duck.	37	42
	PART IV. GEOMETRIC IDENTIFICATION		
39.	Mark each circle in the frame	54	49
36.	Set blocks or things, mark each which is all black	41	62
40.	Mark each triangle	35	36
37.	Mark each block/thing which has some black and some white	28	49
38.	Set different shapes, mark those with more than two sides	14	33
	PART V. SPATIAL JUDGMENT		
53.	Pictures two pieces string, one tight, one loose. Mark the string which you think is longer.	49	20
69.	Two jars same size. Water put in other jar, how high?	46	38
70.	Two jars, one ½ dia. other, mark height water other jar.	34	11
45.	Seriation-sets of lines, make marks missing set (3)	26	35
46.	Seriation-sets of lines, make marks missing set (4)	21	30
54.	Bar graph, fill in or mark to look right	14	29
	PART VI. TIME IDENTIFICATION		
26.	Four frames clocks/watches, mark six o'clock	49	49
25.	Four frames clocks/watches, mark twelve o'clock	45	36
29.	Clock face, made two hands show 6 o'clock	35	42
	PART VII. NUMERAL-NUMBER CONCEPTS		
	The children confronted with numerals 0 through 10 for following items:		
71.	Mark the number for 3	73	47
72.	Mark number for 6	64	52
74.	Mark number that shows number fingers one hand (4 or 5)	64	52
73.	Mark number that shows how many eyes you have	63	51
76.	Mark the number that comes just before 5	52	53
75.	Mark the number which shows you don't have any	47	46
77.	Mark the number that comes just after 7	45	43
78.	Mark the number that comes two numbers before 4	41	27

Test Item	Description of item--reflects instructions read to children	Percent * correct	Point ** Biserial Correlation
79.	Mark the number that comes two numbers after 3	32	22
80.	When I count 5, 4, 3, 2, mark number I should say next	29	38
81.	When I count 10, 9, 8, 7, mark number I should say next	29	37
	PART VIII. MONEY IDENTIFICATION		
48.	Mark the frame that shows a dime	71	24
49.	Mark the frame. . . which of 4 coins will buy the most (3 cents and nickle)	68	33
50.	Mark which will buy the most (2 frames)	56	14
47.	Mark the frame that shows a nickel	54	13
	PART IX. FRACTIONAL RECOGNITION		
32.	Mark the jar that is $\frac{1}{2}$ full of water	41	38
30.	Mark the frame that shows $\frac{1}{4}$ or one-quarter pie	35	17
31.	Mark the frame that shows 1/3 of a pie	24	14
33.	Mark the jar that is $\frac{1}{4}$ full of water	12	01

* Percent correct is computed including errors and omits

**Point biserial correlation is correlation of item with total test score

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

AFTER A SIGNIFICANT F IN ANOVA?

ROBERT R. SMAWLEY

Central Washington State College
Ellensburg

ABSTRACT

From time to time, investigators have proposed various solutions to the problem of identifying all contrasts in classic ANOVA design. A few years ago, Sparks (2) reviewed and then compared four of the newer procedures in their application to a completely randomized 4-method x 7-class design. Specifically, only significant differences between pairs of means were dealt with in the methodological comparisons. In this paper, Tukey's "Compact" (5) and the Link-Wallace (3:119-22) ANOVA shortcuts are also applied, using Sparks' data.

TUKEY'S TEST

TUKEY'S "compact 2-sample test" (5) has been described as the slickest significance check yet devised; experience with it compels one to marvel at its efficiency and simplicity (1). Only a few minutes need to be devoted to various analyses, in contrast with all other methods. Distribution-free in rationale, it *is* a *k*-sample test for location.

Taking a pair of samples at a time, say *a* and *b*, the compact test requires only visual inspection to find which group has the highest value, score or measure. If Group *a* has it, and the lowest value is in Group *b*, then add the number of values in Group *a* exceeding all those in *b*, to the number of values in Group *b* which are below all the values in Group *a*. The critical counts are as follows:

I. For a total count of 7, the 2-tail test equivalence is for 5 percent level significance.

II. For a total count of 10, the 2-tail test equivalence is for 1 percent level significance.

III. For a total count of 13, the 2-tail test equivalence is for the .1 percent level significance.

Sparks Data (2)

The design data used for analysis were mean scores of twenty-eight classes randomly assigned to four groups; the four groups were in turn randomly assigned to the four treatment methods. Rearranged in Table 1, the mean scores are given.

TABLE 1

SPARKS 4-METHOD x 7-CLASS MEAN SCORES

	CLASSES									
	1	2	3	4	5	6	7	$\overline{X}$'s	Sums	Ranges
Method a	14.59	23.44	25.43	18.15	20.82	14.06	14.26	18.68	130.75	11.37
Method b	20.27	26.84	14.71	22.34	19.49	24.92	20.20	21.25	148.77	12.13
Method c	27.82	24.92	28.68	23.32	32.85	33.90	23.42	27.84	194.91	10.58
Method d	33.16	26.93	30.43	36.43	37.04	29.76	33.88	32.52	227.63	10.11
									Sum of Ranges =	44.19

Testing the overall hypothesis that the means are equal ($u1 = u2 = u3 = u4$), an analysis of variance yielded the results in Table 2.

TABLE 2

ANOVA RESULTS FOR SPARKS DATA

Variant Source	df	Sum of Squares	Mean Square	F
Among the Methods	3	830.19	276.73	15.93
Within the Methods	24	416.76	17.37	
Total	27	1246.95		

The tabled F value at the .05 level is 3.01 and the hypothesis was therefore rejected. The question is then begged, "Which treatment means are significantly different?"

It is well known that most authorities are agreed, subsequent individual t-tests by pairs of means should not be performed (4). Sparks (2) also provides an excellent summary of the pitfalls.

Method Comparisons

Sparks' analyses (2:346-8), applying methods contributed by Duncan, Tukey, Scheffé, and Dunn respectively, obtained significant pairs marked with (*) asterisks as given below:

A. Duncan's Multiple Range Test

	d.	c.	b.
a.	*	*	n.s
b.	*	*	
c.	n.s.		

C. Scheffé's S-Method

	d.	c.	b.
a.	*	*	n.s.
b.	*	n.s.	
c.	n.s.		

B. Tukey's (earlier) W-Procedure

	d.	c.	b.
a.	*	*	n.s.
b.	*	*	
c.	n.s.		

D. Dunn's "C" Procedure

	d.	c.	b.
a.	*	*	n.s.
b.	*	*	
c.	n.s.		

Tukey's newer "compact test" (5) produced the results below, similar to his earlier W procedure, Duncan's Multiple Range test, and to Dunn's "c" technique; further, his "compact" method is seen to be more powerful than Scheffé's S method in this case. However, two advantages are evident. Tukey's "compact" determines the level of significance and involves much less labor in application than any of the other methods.

On the other hand, using Duncan's test increases the likelihood of committing a Type I error and is regarded by many as too sensitive where the k-sample problem exceeds four.

E. Tukey's "compact" (5) and Significance Levels of Pairs

	d.	c.	b.
a.	*	*	n.s.
b.	*	*	
c.	n.s.		

	d.	c.	b.
a.	.001	.05	n.s.
b.	.001	.05	
c.	n.s.		

Demonstrating Tukey's newer test and its efficiency with one pair of treatments, a and d above, we find the highest score in row d, the value being 37.04 and the lowest in row a, mean value of 14.06. We count seven values (all) in d higher than the top value in a. We also count all seven means in a lower than the bottom d group value. The sum is fourteen, significant level being .001 as seen in III "critical count" rule for Tukey's "compact test." Likewise, taking the pair b and c, the highest mean is 33.90 in c; the lowest is 14.71 in b treatment. Therefore, the "higher than" count in c is four; the "lower than" in b equals five. The total value count is nine, significant at the .05 level, under rule I above.

Experience from scores of trials with Tukey's "compact" in many different designs supports certain observations. First, samples should be six or more in size, but the n's need not be equal. Second, where ties exist, one can simply cancel out such values, whether highs, lows, or in between. Third, one must comment on the use of significance tests. If it is true the primary aim of a test is to dampen the zeal of the researcher, he is asked to confess that he wants only confirmation of what he already believes. Tukey's "compact" relieves the investigator of choosing the significance level; reduction of bias toward either Type I or Type II errors is obtained. The critical counts and levels are called as they're found, not as one would prefer them, to accept or reject hypotheses.

Link-Wallace ANOVA

For those investigators who prefer parametric procedures in the k-sample problem, perhaps the best bet is the Link and Wallace shortcut to analysis of variance (3:119-122). It is still by far the quickest and most powerful procedure for all ANOVA problems. The method enables the investigator to determine not only whether real differences exist, but also sets "allowances" similar to confidence limits. Further, either one-way or two-way classifications may be dealt with, one observation per cell, with ranges as the variation measure. Experiment-wise error rates are also included for both 5 percent and 1 percent limits. In the present case, the Sparks' class means are "single" observations, one per cell.

From the tabled values calculated by Link and Wallace (3:146) for numbers of groups, and the sample sizes in groups, one first obtains the critical factor value needed. For four groups, of size seven each, the critical factor = .95 and multiplying that value by the sum of all four ranges, (.95) (44.19) = 42 when rounded properly. With the difference limits ("allowances") = 42 as the critical difference between treatment sums, and therefore if any two sums differ by ≥ 42, then one recognizes a real difference, i.e., statistically significant in the usual sense. Readers may compare with previous results.

TABLE 3

LINK-WALLACE METHOD CONFIDENCE INTERVALS

Method Comparisons	Sum Difference (Rounded)	Sums $\bar{X}_{D's}$	Critical 5% -level allowance	95% Confidence Limits on Sums Difference Totals	95% Confidence Limits on Sum Mean Differences
a & b	18	2. 6	±42	-24 to 60	-3. 4 to 8. 6
* a & c	64	9. 2	±42	22 to 106	3. 2 to 15. 2
* a & d	97	13. 8	±42	55 to 139	7. 8 to 19. 8
* b & c	46	6. 6	±42	4 to 88	. 6 to 12. 6
* b & d	79	11. 3	±42	37 to 121	5. 3 to 17. 3
c & d	33	4. 7	±42	-9 to 75	-1. 3 to 10. 7

In the Link-Wallce significance matrix, the pairs marked (*) are all $P \leq .05$ for these same data.

F. Link-Wallace Results

	d.	c.	b.
a.	*	*	(n. s.)
b.	*	*	
c.	(n. s.)		

The last column intervals in Table 3, under F above, are obtained by dividing the 95 percent confidence limits by n=7 "rows" in this example, i. e., column labeled E/n-rows. Statistical consistency is observed for the treatment pair comparisons a & b, and c & d, where the confidence intervals ("allowances") include zero, for a non-significance proof.

SUMMARY

Applications of six analytic methods proposed, after a significant F is obtained in the k-sample randomized design, have been discussed. Results were compared and differences among the methods were observed. Certain interesting similarities were observed.

Summing up, then, we see that the various results obtained with Duncan's, Tukey's earlier "W," Dunn's "c," Tukey's newer "compact," and the Link-Wallace ANOVA shortcut all agree. Scheffé's S was found to be less sensitive, but using it would lower the Type I error risk. Finally, if pressed for time, and many will ask, "Who isn't?", the Tukey newer test for location and the Link-Wallace shortcut-ANOVA analytic procedure will serve the investigator's needs with respectable "practical power," as Tukey has expressed it.

FOOTNOTE

1. Sparks' article (2) includes both original references and detailed steps for the four methods developed by Duncan, Tukey in his earlier W-procedure, Scheffé, and Dunn; hence, that information is omitted in this text and the following references.

REFERENCES

1. Basler, D.D. and Smawley, R.B., "Tukey's Compact vs Classic Test," The Journal of Experimental Education, 36:86-88, Spring 1968.

2. Sparks, J.N., "Expository Notes on the Problem of Making Multiple Comparisons in a Completely Randomized Design," The Journal of Experimental Education, 31:343-349, Summer 1963.

3. Tate, M.W. and Clelland, R.C., Nonparametric and Shortcut Statistics, Interstate Printers and Publishers, Inc., Danville, Illinois, 1957, 171 pp.

4. Taylor, P.A. and Ford, S.F., "Book Reviews," Educational and Psychological Measurement, 27:883, Winter 1967.

5. Tukey, J.W., "A Quick Compact Two-Sample Test to Duckworth's Specifications," Technometrics, 1:31-48, February 1959.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

LOGICAL CONSISTENCY OF STUDENT TEACHERS IN DIFFERENT EDUCATIONAL ENVIRONMENTS

HENRY R. WEINSTOCK
University of Missouri-St. Louis

CHARLES M. PECCOLO and FLOYD L. COPPEDGE
Kansas State University

LLOYD L. COPPEDGE
Northeastern State College

ABSTRACT

Student teachers enrolled at Kansas State University and Northeastern State College (Oklahoma) in the spring of 1967 were compared on (1) maintenance of philosophic orientation, (2) adherence to teaching attitudes, (3) reaction to student-teaching, and (4) performance on related criteria. The GNC Scale of Logical Consistency in Educational Ideas determined their philosophic orientation. The Minnesota Teacher Attitude Inventory determined their attitudes toward potential teaching situations. Comparisons were made before and after student teaching by institution, teaching level, and sex. GNC, MTAI, and ACT levels, methods and student-teaching grades, and grade point averages were tested for intercorrelation.

K. S. U. students were more logically consistent and empirically oriented in their ideas about education before and after teaching participation. Teaching attitudes were best maintained by the elementary students in both institutions. N. S. C. students showed greater intercorrelation among all the criteria. Low logical consistency by N. S. C. students may signify disagreement between theoretical and practical educational views. Maintenance of teaching attitudes by K. S. U. elementary students suggests variable effects of professional courses. Agreement among academic and educational criteria by N. S. C. students suggests a greater unaminity of educational ends and means.

A COMPARISON of the teacher education programs at Kansas State University and Northeastern State College (Oklahoma) was undertaken in the spring of 1967. The research conducted was to determine what, if any, differences existed between and among the student teachers with respect to: (1) the logical consistency with which they held ideas about education prior to teaching participation, (2) the extent to which they held commonly-ascribed attitudes toward teaching prior to teaching participation, (3) the degree with which they had maintained their logical consistency and teaching attitudes upon completion of their relative periods of teaching participation, and (4) the levels they achieved on scholastic criteria pertinent to their respective college programs.

It should be noted that prior research with student teachers at K. S. U. has previously been conducted with regard to logical consistency, teaching attitudes, and related scholastic criteria (5, 4, 6). The results then obtained strongly indicated that inter-institutional research in these areas should be conducted.

PROCEDURE

The student teaching program at Kansas State University required satisfactory performance in a common professional requirement of 20 semester hours of both its secondary and elementary education students. Of these, 6 hours of credit were given for their 7-week actual teaching participation period in the public schools during the senior year. The remainder was earned, prior to student teaching, in educational psychology, and concurrent with student teaching, in educational sociology and either (1) secondary principles and methods, or (2) elementary reading and methods. Supervision of the student teachers was conducted by staff members of the College of Education specializing in the particular teaching fields represented.

The student teaching program at Northeastern

TABLE 1

COMPOSITION OF THE STUDENT TEACHER SAMPLE

Student Teachers Grouped by		Kansas State University			Northeastern State College		
		Number Used	Total Number	Percent* Used	Number Used	Total Number	Percent* Used
School		90	142	60	194	298	65
Level	Secondary	49	77	64	119	187	63
	Elementary	41	65	63	75	112	67
Sex	Male	15	26	58	76	121	63
	Female	75	116	65	118	177	67

*All student teachers on whom the necessary test, professional, and academic data was obtainable are included in the study.

State College required completion of from 21 to 30 hours in its professional program for both secondary and elementary students. Of these, 12 to 16 hours were granted to the secondary and elementary seniors enrolled in the complete semester of full-time intern teaching. The remaining hours consisted of common courses in educational foundations, psychology, methods, and electives. Supervision of intern teaching was conducted by staff members of the Division of Education on a geographical, rather than teaching field basis.

For purposes of this study, the samples of ninety student teachers at Kansas State University and 194 student teachers at Northeastern State College were dichotomized with respect to teaching levels and sex (Table 1). The logical consistency and philosophical orientation with which they held ideas about education was determined by use of the GNC Scale (2) . The GNC Scale utilized a force-sort technique to determine the extent to which the student teacher could separate fifty rationalistic statements (prior to the age of science) from fifty empirical statements (since the advent of science) relating educational theory with practice.

The teaching attitudes of the student teachers at both institutions were determined by the Minnesota Teacher Attitude Inventory (1). This scale was utilized for purposes of assessing each of the groups' attitudes toward the future teaching situations in which they might find themselves.

Comparisons were made between the two institutions on basis of GNC and MTAI scores attained prior to and upon returning from teaching participation. This was done with respect to both the teaching level and the sex of the student teachers. Comparisons on the same criteria were also made within the two schools in order to determine any significant changes inherent within them.

An investigation was then conducted to determine the relationship of student teacher performance regarding the ACT Battery, GNC Scale, MTAI, student teaching participation, cumulative grade point averages, and courses in Principles of Education, Methods of Teaching, and Educational Sociology. They were compared on the basis of their ACT performance both within each school and between the two schools. Lastly, tests for intercorrelation among performance on all criteria by the student teachers were conducted within each of their respective institutions.

PRESENTATION OF DATA

The K. S. U. student teachers, prior to teaching participation, were logically consistent in their ideas about education (Table 2). With the exception of the male group, all GNC means were beyond the . 01 level, the elementary and female groups being the highest. No N. S. C. student teacher groups were found to have attained a significant level of logical consistency ($p \overset{-}{>} .05$). The K. S. U. scores attained prior to student teaching were significantly greater than those of their respective N. S. C. counterparts (Table 3). In both institutions the females were found to be more logically consistent than the males prior to student teaching (Table 4).

With regard to teaching attitudes held prior to teaching participation, significantly higher MTAI scores were attained by the K. S. U. student teachers (Table 3). Although no differences in teaching attitudes were detected at K. S. U. , highly significant differences were found at N. S. C. with regard to teaching level and sex (Table 4).

Upon completion of their respective teaching participation periods, only one shift in logical consistency between the student teachers of the two schools

TABLE 2

LOGICAL CONSISTENCY OF STUDENT TEACHERS, AS MEASURED BY THE GNC SCALE

Student Teachers Grouped by		Before Student Teaching		After Student Teaching	
		K. S. U. mean	N. S. C. mean	K. S. U. mean	N. S. C. mean
School		47***	29	46***	30
Level	Secondary	45***	30	46***	30
	Elementary	49***	28	46***	29
Sex	Male	38**	27	38**	29
	Female	49***	31	48***	30

*significant at .05 level (GNC = ±32)
**significant at .02 level (GNC = ±37)
***significant at .01 level (GNC = ±41)

was noted. The males no longer differed on this measure, although prior to student teaching the K. S. U. males had been higher (Table 3). All of the K. S. U. groups, however, were still at the .02 or higher GNC level, whereas each of the N. S. C. groups was still insignificant (Table 2). In the maintenance of desirable teaching attitudes, the elementary levels and females in both schools showed significantly higher scores both before and after student teaching (Table 4).

With regard to intra-group comparisons, a consistent pattern was apparent. Prior to student teaching, the K. S. U. females had been significantly higher than the males in logical consistency; the same situation held upon completion of this period. However, a large increase in logically consistent males and a corresponding decrease in logically consistent female student teachers at N. S. C. was noted.

A comparison of ACT Battery performance on behalf of the two schools also revealed highly significant differences to exist between the two schools (Table 5). In each instance the K. S. U. student teachers had achieved significantly higher scores on this test. The smallest of these was found in the comparison of the males in each school. Within each school itself, however, the only significantly higher scores on the ACT Battery were those of the N. S. C. secondary student teachers as compared to their elementary counterparts.

Several highly significant intercorrelations among the test, professional, and scholastic criteria within each school were detected. Among the N. S. C. student teachers, a total of forty pair of criteria correlated significantly (Table 7); however, only twenty-six were found for the K. S. U. student teachers (Table 6). Among the K. S. U. students only Methods correlated with Student Teaching, whereas it correlated with all nine at N. S. C. The identical relationship was found to exist with regard to the GPA criterion measure for each institution. Lastly, it was found that the first and second largest correlation coefficient in each school was, respectively, the score achieved among the MTAI and GNC tests taken before and after teaching participation.

CONCLUSIONS

A comparison of the spring 1967 student teachers at Kansas State University and Northeastern State College showed significant differences to exist in the logical consistency with which ideas about education were held by their respective student teachers. It had been earlier hypothesized, however, that exposure to rationalistic views dealing with the methods of educating man, on the other hand, results in conflicting frameworks of educational reference on the part of the teacher (4). Thus the educational environment evidencing significantly lower proportions and levels of logically consistent student teachers may be indicative of either (1) a lesser empirical stress in its general educational program, (2) a more severe conflict between students' views of mankind from theoretical and practical standpoints, or perhaps (3) a willingness to accept different points of view.

Significant differences were detected in desirable attitudes displayed toward their future teaching tasks by the student teachers. These differences, however, also existed in the same order within each institution, i. e., the secondary level was lower than the elementary and the males were lower than the females. Most of the elementary student teachers at both schools were females, however (K. S. U. 98 percent, N. S. C. 84 percent). As lower institutional teaching attitudes thus tend to reflect specific differences between the female students at each campus, significantly higher teaching attitude test scores may indicate that (1) female student teachers are more affected by their professional courses than are males,

TABLE 3

COMPARISON OF LOGICAL CONSISTENCY AND TEACHING ATTITUDES BETWEEN SCHOOLS

Student Teachers Compared by		Before Student Teaching		After Student Teaching	
		GNC t	MTAI t	GNC t	MTAI t
School	K. S. U. and N. S. C.	9.10**	7.71**	7.98**	6.33**
Level	K. S. U. Secondary and N. S. C. Secondary	5.50**	5.82**	5.93**	4.56**
	K. S. U. Elementary and N. S. C. Elementary	7.79**	5.02**	5.30**	4.23**
Sex	K. S. U. Males and N. S. C. Males	2.26*	2.85**	1.87	1.88
	K. S. U. Females and N. S. C. Females	8.64**	6.40**	7.57**	5.37**

*significant at .05 level
**significant at .01 level

TABLE 4

COMPARISON OF LOGICAL CONSISTENCY AND TEACHING ATTITUDES WITHIN SCHOOLS

Student Teachers Compared Within		Before Student Teaching		After Student Teaching	
		GNC t	MTAI t	GNC t	MTAI t
Level	K. S. U. Secondary and K. S. U. Elementary	-1.47	-2.01*	0.16	-2.12*
	N. S. C. Secondary and N. S. C. Elementary	0.88	-3.06**	0.21	-2.01*
Sex	K. S. U. Males and K. S. U. Females	-2.78**	-3.00**	-2.81**	-3.14**
	N. S. C. Males and N. S. C. Females	-1.67	-4.06**	-0.43	-3.13**

*significant at .05 level
**significant at .01 level

(2) more professional courses, as required for elementary preparation, are more effective in shaping teacher attitudes, or (3) the effectiveness of professional courses for teachers may vary institutionally.

With regard to differences in the educational environments, the type and length of teaching participation seemed to have little effect on the maintenance of either logical consistency in educational thought or predetermined attitudes toward teaching. Consistently smaller differences (smaller t - values) with respect to school, level, and sex on these two measures seemed to indicate a possible reapproachment between the two institutions. These diminishing differences may thus be the result of (1) the longer, one-semester student teaching period required of N. S. C. students, as contrasted to the half-semester period at K. S. U., (2) different challenges to the prior-developed educational ideas and attitudes of the student teachers, (3) differences in the effect of long-term student teacher supervision done on either a level-basis (independent of geographical restrictions, as at K. S. U.) or a geographical basis (independent of level restrictions, as at N. S. C.).

With respect to relationships between and among performances on pertinent scholastic criteria, the student teachers at the two institutions showed contrasting profiles. Although the higher ACT scores of the K. S. U. students barely related to their higher GNC and MTAI achievement, the lower ACT scores of the N. S. C. students were, however, highly related to both of these tests. This may thus indicate that the ACT Battery of college achievement predictability neither measures logical consistency nor teaching attitudes very well.

From an intra-school standpoint, the larger number of significant correlations at N. S. C. is accounted for primarily by more consistent relationships of Methods and Student Teaching grades with all other criteria. On the other hand, the total lack of relationship in both schools between Student Teaching grades and GNC-MTAI scores may indicate some communality in the student teacher environments. This seems to be supported by the similar correlations among GNC-MTAI scores achieved by both groups prior to their teaching participation. It is, of course, also possible that the GNC and MTAI may not be significant predictors of success in the completion of a teacher education program at either institution.

Lastly, more numerous and/or higher relationships between general and professional scholastic criteria as noted at N. S. C. may be indicative of a greater campus-wide unanimity of efforts toward the specific goal of teacher-preparation. Consequently, it seems worthwhile to further investigate the relationship which exists between the general and professional educational elements of the multi-university, on the one hand, and the teacher-education oriented college, on the other.

TABLE 5

COMPARISON OF ACT BATTERY SCORES OF STUDENT TEACHERS BETWEEN AND WITHIN SCHOOLS

Student Teachers Compared by	Between Schools	t	Within Schools	t
School	K. S. U. and N. S. C.	9. 15**		—
Level	K. S. U. Secondary and N. S. C. Secondary	6. 43**	K. S. U. Secondary and K. S. U. Elementary	1. 57
	K. S. U. Elementary and N. S. C. Elementary	7. 45**	N. S. C. Secondary and N. S. C. Elementary	4. 29**
Sex	K. S. U. Males and N. S. C. Males	2. 73**	K. S. U. Males and K. S. U. Females	-0. 86
	K. S. U. Females and N. S. C. Females	9. 27**	N. S. C. Males and N. S. C. Females	0. 98

*significant at . 05 level
**significant at . 01 level

TABLE 6

INTERCORRELATIONS OF CRITERIA FOR STUDENT TEACHERS AT K. S. U. (N = 90)

		ACT 1	GNC(1) 2	GNC(2) 3	MTAI(1) 4	MTAI(2) 5	Princ 6	Meth 7	Stu Tch 8	Ed Soc 9	GPA 10
ACT	1		.232*	.274*	.182	.029	.349**	.385**	-.040	.236*	.455**
GNC(1)	2			.697**	.372**	.305**	.289**	.012	-.144	.217*	.166
GNC(2)	3				.338**	.354**	.304**	.087	-.008	.308**	.287**
MTAI(1)	4					.725**	.223*	-.089	-.043	.227*	.119
MTAI(2)	5						.133	-.215	-.015	.149	-.015
Princ	6							.149	-.052	.271**	.372**
Meth	7								.263*	.218*	.426**
Stu Tch	8									.153	.188
Ed Soc	9										.295**
GPA	10										

*significant at .05 level
**significant at .01 level

TABLE 7

INTERCORRELATIONS OF CRITERIA FOR STUDENT TEACHERS AT N. S. C. (N=194)

		ACT 1	GNC(1) 2	GNC(2) 3	MTAI(1) 4	MTAI (2) 5	Princ 6	Meth 7	Stu Tch 8	Ed Soc 9	GPA 10
ACT	1		.460**	.387**	.197**	.144*	.297**	.240**	.199**	.314**	.412**
GNC(1)	2			.685**	.429**	.367**	.326**	.270**	.111	.310**	.446**
GNC(2)	3				.436**	.402**	.286**	.181*	.107	.258**	.318**
MTAI(1)	4					.791**	.232**	.347**	.071	.164*	.277**
MTAI(2)	5						.208**	.266**	.114	.141	.220**
Princ	6							.334**	.286**	.408**	.562**
Meth	7								.245**	.287**	.428**
Stu Tch	8									.286**	.305**
Ed Soc	9										.506**
GPA	10										

*significant at .05 level
**significant at .01 level

REFERENCES

1. Cook, W. W., Leeds, C. H., and Callis, R., Minnesota Teacher Attitude Inventory, Psychological Corporation, New York, 1951.

2. Gowin, D. B., Newsome, G. L., and Chandler, K. A., ''A Scale to Study Logical Consistency of Ideas About Education,'' Journal of Psychology, 51: 443-445, 1961.

3. Weinstock, H. R., ''Logical Consistency And Self-ascribed Metaphysical, Epistemological, and Ethical Views of Teachers Enrolled in Philosophy of Education,'' 18th Annual Proceedings of the Southwestern Philosophy of Education Society, 18:34-42, 1968.

4. Weinstock, H. R., and Peccolo, C. M.,''Logical Consistency in Ideas About Teaching Held By Student Teachers,'' A. E. R. A. Paper Abstracts 1968.

5. Weinstock, H. R., and Peccolo, C. M.,''Logical Consistency of Secondary Student Teachers and Other Psychological, Academic, and Curricular Aspects of Teacher Education at Kansas State University,'' Journal of Teacher Education, 17: 498-499, 1966.

6. Weinstock, H. R. and Peccolo, C. M., ''Secondary and Elementary Student Teachers Compared on Logical Consistency and Teaching Attitudes,'' The Elementary School Journal (in press).

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

BIOGRAPHICAL FACTORS AND TEACHER EFFECTIVENESS

JOHN A. WILLIAMSON

North Texas State University

ABSTRACT

This study attempted to detemine whether biographical factors might be identified which would distinguish between teachers receiving high and low student assessments. More than 3,000 high school seniors named the most and least effective teacher they had experienced. Most and least effective were defined in terms of the three major patterns of classroom teacher behavior identified by Ryans. Teachers named by three or more seniors completed a biographical questionnaire. Effective teachers numbering 125 and ineffective teachers numbering 81 responded. A coefficient of correlation was found for each of the biographical items and the criterion, teacher rating. Significant correlations were found for amount of mother's formal education, mobility of teacher, number of books read annually, home ownership, and student-teacher relationships.

THE SUCCESS or failure of any school's educational endeavor rests largely with the classroom teacher, and there is no substitute for an effective one. However, the identification of the effective teacher has been a problem that has plagued educational leaders for years, and it has yet to be resolved. Despite the voluminous research that has been devoted to teacher effectiveness, results have been inconclusive and inconsistent. As Biddle (2) has pointed out, "Few, if any, facts are now deemed established about teacher effectiveness, and many former 'findings' have been repudiated." However, few, if any, problems are more crucial in education and, despite the paucity of definitive findings, efforts to understand teacher effectiveness must continue.

The purpose of this study was to determine whether biographical factors may be identified which would distinguish between teachers receiving high and low student assessments.

Of major concern in any study of teacher effectiveness is the method of evaluation. Several techniques have been used in previous research, including pupil achievement, ratings by various observers, standardized instruments, and many others. However, because the student is the one person who observes the teacher in his day-to-day teaching activities and, consequently, is in the best position to evaluate the effectiveness of that teacher, student evaluation was used in this study.

PROCEDURES

A simple rating instrument was administered to more than 3,000 Colorado high school seniors at their respective schools. The seniors were asked to consider every teacher whom they had encountered throughout their school experience and then to identify by name the one teacher who was the most effective teacher they had encountered. They were also asked to identify the least effective teacher they had experienced. "Most effective" and "least effective" were defined in terms of the three major patterns of teacher classroom behavior identified by Ryans (4). The three dimensions are:

1.	friendly, understanding behavior	versus	aloof, egocentric, restricted behavior
2.	responsible, businesslike, systematic behavior	versus	unplanned, slipshod behavior
3.	stimulating, imaginative , enthusiastic behavior	versus	dull, routine behavior

A total of 669 most effective teachers and 505 least effective teachers were named by the seniors. ("Most" and "least" effective will hereafter be referred to as effective and ineffective, respectively.) Of this number, 309 teachers were named by three or more seniors. Of these 309 teachers, 281 were located and were asked by mail to complete a biographical questionnaire. The questionnaire was designed to gain some insight into the background of the teacher and the opinions of the teacher regarding various aspects of teacher-pupil relationships. Of the 281 teachers, 125 of 149 effective and 81 of 132 ineffective teachers responded. This represented a response of 84 percent for the effective teachers and 61 percent for the ineffective teachers.

In an attempt to utilize only those biographical items which seemed most appropriate to the criterion, only twenty items in the questionnaire were included in this study. A coefficient of correlation was found for each of the items and the criterion, teacher rating.

FINDINGS AND DISCUSSION

Significant correlations were found for five of the items (see Table 1). The items were: amount of mother's formal education, status of home ownership, number of books read for pleasure each year, number of times residence had changed during last five years, and an opinion item regarding the teacher's attitude toward sending a pupil to the office for discipline.

TABLE 1

CORRELATIONS BETWEEN TEACHER RATING AND TWENTY VARIABLES

Variable	r
1. Teacher sex	0.01
2. Number of children	-0.04
3. Teacher age	0.02
4. Total teaching experience	0.13
5. Mother's education	0.18**
6. Father's education	0.01
7. Number of siblings	-0.03
8. Order of birth	0.11
9. Size of home city	0.05
10. Decision to teach	0.06
11. Distance home was from college	-0.03
12. Father's occupation	0.05
13. Spouse's education	0.03
14. Home ownership	-0.15*
15. Send student to office for discipline	0.24**
16. Teachers do tasks willingly	-0.04
17. Pupils are competent to select discussion topics	-0.05
18. Pupils try to do their best	0.06
19. Books read per year	0.22**
20. Change of residence during last five years	-0.18**

* Significant at the .05 level.

** Significant at the .01 level.

Results reported in Table 1 indicate that certain data from a biographical questionnaire correlate with teacher effectiveness. However, the correlations were rather low; consequently, from the standpoint of practical significance, they should be regarded as only trends. The findings indicated that effective teachers were differentiated from ineffective teachers in five ways: (1) effective teachers read more books for pleasure than did ineffective teachers, (2) the mothers of effective teachers had more formal education than did mothers of ineffective teachers, (3) effective teachers changed their residence less often than did ineffective teachers, (4) effective teachers tended to own their homes more than ineffective teachers, and (5) effective teachers were inclined to blame themselves in the majority of cases when they sent a student to the office for discipline, whereas ineffective teachers placed most of the blame on the student.

The apparent mobility of ineffective teachers as indicated by two of the findings is not surprising. That ineffective teachers would tend to move more frequently than effective teachers may be attributed to at least two factors: (1) The fact that students recognize the teacher's ineffectiveness should indicate that the school principal may also be aware of this and, consequently, this could result in the dismissal of the teacher. This is a relevant factor despite the findings of Barr and Leeds that principals and students judge behavior of teachers differently (1, 3). (2) Ineffective teachers often experience poor teacher-pupil relationships which may create unpleasant and, sometimes, intolerable conditions for the teacher. As a consequence, in an attempt to escape this undesirable situation, the teacher may seek employment in another locale.

That the effective teachers tended to blame themselves in the majority of cases when they sent a student to the office for discipline appears to indicate that the effective teachers are more aware of the feelings of the students than the ineffective teachers. It would also seem to be indicative of a more harmonious relationship between the effective teachers and their students than the relationships existing between the ineffective teachers and their students.

SUMMARY

This study attempted to determine if relationships existed between student ratings of teachers and biographical characteristics of teachers. Several relationships were found. Mothers of effective teachers had more formal education than mothers of ineffective teachers; effective teachers were less mobile than ineffective teachers, and effective teachers experienced more harmonious relationships with students than did ineffective teachers.

APPENDIX A

Questionnaire

DIRECTIONS: Place the appropriate number in the blank provided to the left of the question.

____ 1. Sex
1. Male
0. Female

____ 2. Number of children, if applicable:
1. None
2. One
3. Two
4. Three
5. Four
6. Five
7. Six
8. Seven or more

____ 3. My age is
1. 20-24
2. 25-29
3. 30-34
4. 35-39
5. 40-44
6. 45-49
7. 50-54
8. 55-59
9. 60-64
10. 65 or over

____ 4. Total number of years teaching experience excluding this year.

____ 5. My mother's formal education was that of:
1. Eighth grade or less
2. High school, but not a graduate
3. High school graduate
4. College, but not a graduate
5. College graduate

____ 6. My father's formal education was that of:
1. Eighth grade or less
2. High school, but not a graduate
3. High school graduate
4. College, but not a graduate
5. College graduate

____ 7. The number of brothers and sisters I have is: (Include living and deceased.)
1. Zero
2. One
3. Two-Three
4. Four-five
5. Six or more

____ 8. My order of birth in my family was:
1. First
2. Second
3. Third
4. Fourth
5. Fifth or beyond

____ 9. During most of my childhood I lived in:
1. The country
2. A very small town (under 1,000)
3. A small town (1,000-5,000)
4. A town (5,000-10,000)
5. A small city (10,000-50,000)
6. A city (50,000-100,000)
7. A large city (over 100,000)

____ 10. As you recall, was your decision to be a teacher made:
1. Prior to or during high school
2. After high school, but prior to college
3. During college
4. After college graduation

____ 11. How far was your parental home from the institution which granted your Bachelor's degree ?
1. Within 20 miles
2. 21-100 miles
3. 101-350 miles
4. 351-1,000 miles
5. More than 1,000 miles

____ 12. During most of your childhood and adolescent life, what was the general nature of your father's or guardian's occupation?
1. Professional
2. Farming
3. Skilled (machinist, printer, etc.)
4. Unskilled labor
5. Business

____ 13. How does your spouse's educational level compare with you own? (Answer only if applicable.)
1. My educational level is considerably lower.
2. My educational level is somewhat lower
3. My educational level is about the same
4. My educational level is somewhat higher
5. My educational level is considerably higher

____ 14. What is the status of your home ownership (home, duplex, apartment, etc.) ?
1. Home is owned and fully paid for
2. Home is owned and nearly paid for
3. Home is owned and about half paid for
4. Home is owned and less than half paid for
5. Home is being rented

____ 15. When a teacher sends a pupil to the office, what percentage of the cases do you feel are the fault of the teacher rather than the fault of the pupil? (Indicate one.)
1. Less than 10 percent of the cases
2. 10-20 percent of the cases
3. 20-30 percent of the cases
4. 30-60 percent of the cases

____ 16. In light of your acquaintance with teachers, what do you think of the following statement?
"Most teachers are willing to assume their share of the unpleasant tasks associated with teaching."
1. Strongly disagree
2. Disagree
3. Uncertain
4. Agree
5. Strongly agree

____17. What do you think of the following statement?
"Pupils are usually quite competent to select their own topics for themes and speeches."
1. Strongly disagree
2. Disagree
3. Uncertain
4. Agree
5. Strongly agree

____18. In light of your experience with children, what do you think of the following statement?
"Most pupils try to do their work to the best of their ability."
1. Strongly disagree
2. Disagree
3. Uncertain
4. Agree
5. Strongly agree

____19. How many books do you read for your own pleasure each year?
1. Less than two books
2. 2-5 books
3. 6-10 books
4. More than 10 books

____20. How many times has your residence changed during the last five years? (Do not include moves within the same city.)
1. None 4. Three 7. Six or more
2. One 5. Four
3. Two 6. Five

REFERENCES

1. Barr, Arvil S., "Summary and Comments," The Journal of Experimental Education, 15: 99-100, September 1946.

2. Biddle, Bruce J., and Ellena, William J., ed. Contemporary Research on Teacher Effectiveness, Holt, Rinehart and Winston, New York, 1964, 352 pp.

3. Leeds, Carroll H., "A Second Validity Study of the Minnesota Teacher Attitude Inventory," The Elementary School Journal, 52:398-405, March 1952.

4. Ryans, David G., Characteristics of Teachers, American Council on Education, Washington, D.C., 1960, 416 pp.

Note: This study was conducted while the author was employed at Colorado State College.

From DERS

THE HUMANIST TRADITION IN MODERN INDIAN EDUCATIONAL THOUGHT

By K. G. Sayaidain. With a foreword by John Guy Fowlkes. 1967

256 pages, $5.95

Presents some of the significant contributions made to educational thought by Tagore, Gandhi, Iqbal, Azad, Radhakrishnan, and Husain. Gives Western leaders an insight into Indian educational trends.

"Here we have a very able Indian scholar writing about Indian leaders as an Indian sees them functioning in India and on the international scene."—Prof. CLIFFORD S. LIDDLE, Korea.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

THE QUALITY OF WRITTEN EXPRESSION OF THIRD-GRADE CHILDREN UNDER DIFFERING TIME LIMITS

MARY JO WOODFIN

California State College at Long Beach

ABSTRACT

The purpose of this study was to determine whether significant differences in the quality of the written expression of third-grade children would occur under differing time conditions. Three groups of third-grade children wrote compositions on topics of their own choosing once every 2 weeks for a total of five experimental sessions. One group wrote for 20 minutes; and one for any amount of time from zero to 225. The 225-minute group rated significantly higher than the other two groups in all five trials; the 90-minute group significantly higher than the 20-minute group in all but Trial 2.

TO ESTABLISH time priorities for teaching written expression, one of the questions educators may well ask themselves is what time allotments result in the highest quality of writing. Research evidence concerning the effects of time on the quality of written expression is scarce. One example is the work of Wilson (7) who found that third-grade children wrote longer compositions, on the average, if they wrote for 40-minute periods than if they wrote for 25-minute periods. No attempt was made to investigate the quality of this writing. Many suggestions are offered in the literature by individuals with extensive experience in teaching composition. Some recommend long periods of time (1, 2, 3, 4, 6); some short periods (8).

To find whether the quality of written expression of third-grade children is significantly different under differing time conditions, a study of the quality of this writing under three time procedures was made: 1) Children in the first group could write for any amount of time they wished, from zero to 225 minutes; 2) For the second condition, children could choose any period of time they wished, from zero to a maximum of 90 minutes; and 3) The third time allotment was the customary classroom - one where all children wrote for 20 minutes. In all three time conditions, children were allowed to write on any topic they chose, and motivation time was not included.

The sample was drawn from a large Southern California metropolitan school district. Children in the three groups did not differ significantly from each other in chronological age, IQ, socioeconomic rank, average language test score, or initial quality of writing as measured by the McClellan Scale described below. Initially, the 90 minute group scored significantly higher in reading than the 225 minute group; this difference was significant at the .05 level.

Five trials were held 2 weeks apart to see whether the relationship between time and quality was consistent over a period of weeks. Each group wrote under the same time conditions throughout the study.

A scale developed by McClellan (5) was used to judge the quality of the writing. This scale consists of nine rankings, with one the lowest and nine the highest, and rates imagination, expression of ideas, and other creative elements.

Three judges rated each composition. The average of their three scores was used for all statistical procedures. The average correlation among these judges was .79.

Average ratings on the McClellen Scale for the three time conditions are given in Table 1. Analysis of variance and Duncan Test procedures showed the

TABLE 1

MEANS AND STANDARD DEVIATIONS OF RATINGS ON McCLELLAN SCALE FOR FIVE TRIALS

	225-Minute Group			90-Minute Group			20-Minute Group		
Trial	Number of Children	Mean Rating	Standard Deviation	Number of Children	Mean Rating	Standard Deviation	Number of Children	Mean Rating	Standard Deviation
1	179	4.65	1.79	179	4.13	1.61	184	3.54	1.46
2	184	5.12	1.74	181	4.30	1.56	183	4.01	1.48
3	187	4.92	1.68	180	4.54	1.41	186	3.76	1.40
4	181	5.20	1.71	187	4.82	1.50	185	4.11	1.35
5	174	5.45	1.52	179	4.93	1.37	178	4.08	1.30

following significant differences among these means:

1. The 225-minute group rated significantly higher than the 90-minute and the 20-minute groups in all five trials.

2. The 90-minute group rated significantly higher than the 20-minute group in Trials 1, 3, 4, and 5. All differences were significant at the .01 level except the difference in Trial 4 between the 225- and the 90-minute groups, which was significant at the .05 level.

IMPLICATIONS

If the primary aim of including written expression as a part of the third-grade curriculum is to secure the highest quality in that writing, then the findings from this study would indicate that the best way to obtain this quality is to give children a time choice within long periods of time rather than to require all children to write for a uniform short period. Quality of writing, as measured by the McClellan Scale at least, increases for third-graders with each increase in time available to the children for writing.

The implications for individualizations of instruction in written expression along choice patterns seem clear. And that allowing a choice of time for writing over a period of time continues to result in better writing, as measured by this scale, suggests further implications for classroom scheduling.

REFERENCES

1. Creed, Esther D., "What'll We Write About?" Elementary English, 33:24-28, January 1956.

2. Cross, Alma, "Books and Children's Creative Expression," Elementary English, 35:38-41, January 1958.

3. Diederich, Paul B., "The Measurement of Skill in Writing," School Review, 54:584-592, December 1946.

4. Hardy, Hilda, "The Child, A Creator," Elementary English, 38:491-493, November 1961.

5. McClellan, Jack, "Creative Writing Characteristics of Children," unpublished EdD. dissertation, University of Southern California, 1956, pp. 85-95.

6. Stegall, Carrie, "Take a Number From One to Ten," Elementary English, 39:33-40, 49, January 1962.

7. Wilson, Louis Ada, "A Study of Some Influencing Factors Upon and the Nature of Young Children's Written Language," The Journal of Experimental Education, 31:4:371-380, Summer 1963.

8. Yocum, Charles G., "Creative Writing," Grade Teacher, 81:52-53+, October 1963.

TABLE 2

ANALYSIS OF VARIANCE OF McCLELLAN SCALE I FOR FIVE TRIALS

	Analysis of Variance														
	Trial 1 N = 542			Trial 2 N = 548			Trial 3 N = 553			Trial 4 N = 553			Trial 5 N = 531		
Source of Variance	MS	df	F	MS	df	F	MS	df	F	MS	df	F	MS	df	F
Time (20, 90, 225 minutes)	9.87	2	3.41*	10.59	2	3.75*	13.30	2	5.22**	16.80	2	6.51**	37.54	2	16.23**
IQ (Above and below 105)	113.22	1	39.11**	89.77	1	31.80**	148.14	1	58.16**	103.97	1	40.30**	80.01	1	34.59**
Sex	1.40	1	0.48	3.78	1	1.34	5.55	1	2.18	7.33	1	2.84	11.03	1	4.77*
Time - IQ	55.58	2	19.20**	63.01	2	22.32**	68.62	2	26.94**	50.83	2	19.70**	76.40	2	33.03**
Sex - IQ	3.39	1	1.17	0.20	1	0.07	0.69	1	0.27	5.29	1	2.05	0.12	1	0.05
Time - Sex	2.78	2	0.96	7.54	2	2.67	9.42	2	3.70*	8.05	2	3.12*	5.62	2	2.43
Time - Sex - IQ	2.78	2	0.96	4.63	2	1.64	20.45	2	8.03**	2.43	2	0.94	75.17	2	3.25*
All Groups	23.88	11	8.25**	24.98	11	8.85**	34.84	11	13.68**	26.08	11	10.11**	31.34	11	13.55**
Within Groups	2.90	530	----	2.82	536	----	2.55	541	----	2.58	541	----	2.31	519	----

* Significant at the .05 level
** Significant at the .01 level

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 3, Spring 1969)

TEST RELIABILITY AND PARAMETERS OF OBSERVED SCORE DISTRIBUTIONS[1]

Donald W. Zimmerman

Carleton University
Ottawa, Canada

ABSTRACT

A method of determining the reliability coefficient of a test from a formulation which does not employ the concepts of true score and error score, together with assumptions about the process which generates variability in scores, is described. Several well-known reliability formulas as well as some new results are derived from models which hypothesize different sources of variability in scores.

THE PURPOSE of this paper is to describe a method for constructing models based on assumptions about the nature of variability in test scores. Novick (7) and Lord and Novick (6) characterized a strong model as one in which assumptions are made about the functional form of the distribution of errors. In the present paper the concepts of true score and error score are not employed, and attention is focused on the distribution of observed scores over repeated measurements. For a given person that distribution has the same form as the distribution of error scores, as defined in the classical test theory model. If $X = A + B$, where A is defined as the expected value of X, then $\sigma^2 X = \sigma^2 (EX + B) = \sigma^2 B$. With respect to the variation in a given person's score over repeated measurements, assumptions concerning $\sigma^2 B$ are the same as assumptions concerning $\sigma^2 X$. In the present context a strong model is regarded as one in which the variability of observed scores is specified in terms of parameters such as the number of items.

Several different models are considered. The approach taken in this paper leads to a formulation which unifies some results given by other investigators and indicates the relation among several special reliability coefficients.

DEFINITIONS

The subscript g will refer to repeated measurements on a given person, and the subscript j will refer to persons. Expectations, variances, and covariances will be denoted by E, σ^2, and COV, with subscripts indicating whether these values are taken over repeated measurements, persons, or both. The summation sign is not used.

We consider a set of pairs of independent, repeated measurements having the same distribution function, one member denoted by X_{gj}, the other by X'_{gj}. The quantity X_{gj} is a random variable having expectation over repeated measurements $\underset{g}{E} X_{gj}$ and variance over repeated measurements $\underset{g}{\sigma^2} X_{gj}$. The expectation of X_{gj} over both persons and repeated measurements is $\underset{g,j}{E} X_{gj}$, and the variance of X_{gj} over both persons and repeated measurements is $\underset{g,j}{\sigma^2} X_{gj}$. The distribution functions of X_{gj} and X'_{gj} are the same for any person j, such that $\underset{g}{E} X_{gj} = \underset{g}{E} X'_{gj}$ and $\underset{g}{\sigma^2} X_{gj} = \underset{g}{\sigma^2} X'_{gj}$. It is not assumed that these distribution functions are the same for different persons.

Test reliability is defined as follows:

$$\rho = \frac{\underset{g,j}{COV}(X_{gj}, X'_{gj})}{\underset{g,j}{\sigma} X_{gj} \, \underset{g,j}{\sigma} X'_{gj}},$$

where

$$\underset{g,j}{COV}(X_{gj}, X'_{gj}) = \underset{jg}{EE} X_{gj} X'_{gj} - \underset{jg}{EE} X_{gj} \, \underset{jg}{EE} X'_{gj} .$$

Since repeated measurements are independent,

$$\underset{g}{E}X_{gj}X'_{gj} = \underset{g}{E}X_{gj}\underset{g}{E}X'_{gj} \quad \text{and}$$

$$\underset{g,j}{\mathrm{COV}}(X_{gj}, X'_{gj}) = \underset{j}{E}(\underset{g}{E}X_{gj}\underset{g}{E}X'_{gj}) - \underset{jg}{EE}X_{gj}\underset{jg}{EE}X'_{gj} =$$

$$\underset{j}{\mathrm{COV}}(\underset{g}{E}X_{gj}, \underset{g}{E}X'_{gj}) = \underset{j}{\sigma^2}\underset{g}{E}X_{gj}.$$

Also,

$$\underset{g,j}{\sigma^2}X_{gj} = \underset{jg}{EE}X^2_{gj} - (\underset{jg}{EE}X_{gj})^2 ,$$

$$\underset{g}{\sigma^2}X_{gj} = \underset{g}{E}X^2_{gj} - (\underset{g}{E}X_{gj})^2 ,$$

$$\underset{j}{E}\underset{g}{\sigma^2}X_{gj} = \underset{jg}{EE}X^2_{gj} - \underset{j}{E}(\underset{g}{E}X_{gj})^2 ,$$

and

$$\underset{j}{\sigma^2}\underset{g}{E}X_{gj} = \underset{j}{E}(\underset{g}{E}X_{gj})^2 - (\underset{jg}{EE}X_{gj})^2 .$$

Therefore,

$$(1) \quad \underset{g,j}{\sigma^2}X_{gj} = \underset{jg}{E\sigma^2}X_{gj} + \underset{j}{\sigma^2}\underset{g}{E}X_{gj} .$$

Substituting $\underset{j}{\sigma^2}\underset{g}{E}X_{gj}$ for $\underset{g,j}{\mathrm{COV}}(X_{gj}, X'_{gj})$ in the definition of ρ, using the relationship $\underset{g,j}{\sigma}X_{gj} = \underset{g,j}{\sigma}X'_{gj}$, and using equation (1) gives

$$(2) \quad \rho = \frac{\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{\underset{g,j}{\sigma^2}X_{gj}} = 1 - \frac{\underset{jg}{E\sigma^2}X_{gj}}{\underset{g,j}{\sigma^2}X_{gj}} .$$

Construction of a strong model, as considered in this paper, consists of the following steps. A process which generates variability in scores is hypothesized, and the distribution function of the observed score random variable, X_{gj}, is determined. That function contains parameters such as the number of test items or the probability of chance success. The value of the quantity $\underset{jg}{E\sigma^2}X_{gj}$ is then found, and, by substituting in (1) and (2), equations for the total observed variance and the reliability coefficient are obtained. A reliability coefficient found by this method is specific to the particular source of variability in scores which is hypothesized.

Item Sampling Model: A Coefficient of Equivalence

We consider first the model presented by Lord (5), in which variability in scores is attributed to sampling of items from a population. Each measurement consists of drawing a random sample of N items from a population of items, a certain proportion of which, P_j, are known by person j, and the random variable X_{gj} represents the number of items in the sample known by person j. Over repeated measurements X_{gj} is distributed binomially with $\underset{g}{\sigma^2}X_{gj} = NP_j(1 - P_j)$ and $\underset{g}{E}X_{gj} = NP_j$. Therefore,

$$(Q1) \quad \underset{g}{\sigma^2}X_{gj} = \frac{\underset{g}{E}X_{gj}(N - \underset{g}{E}X_{gj})}{N} .$$

The arithmetic mean of Q1 over persons is

$$\underset{jg}{E\sigma^2}X_{gj} = \frac{N\underset{jg}{EE}X_{gj} - \underset{j}{E}(\underset{g}{E}X_{gj})^2}{N} .$$

Since $\underset{j}{\sigma^2}\underset{g}{E}X_{gj} = \underset{j}{E}(\underset{g}{E}X_{gj})^2 - (\underset{jg}{EE}X_{gj})^2$,

$$\underset{jg}{E\sigma^2}X_{gj} = \frac{\underset{g,j}{E}X_{gj}(N - \underset{g,j}{E}X_{gj})}{N} - \frac{\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{N} .$$

Substituting this result in (1) gives

$$(Q2) \quad \underset{g,j}{\sigma^2}X_{gj} = \frac{(N-1)\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{N} + \frac{\underset{g,j}{E}X_{gj}(N - \underset{g,j}{E}X_{gj})}{N} .$$

Substituting the same result for $\underset{jg}{E\sigma^2}X_{gj}$ in (2), noting that

$$\frac{\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{N\underset{g,j}{\sigma^2}X_{gj}} = \frac{\rho}{N}, \quad \text{and solving for } \rho, \text{ gives}$$

$$(Q3) \quad \rho_Q = \frac{N}{N-1}\left[1 - \frac{\underset{g,j}{E}X_{gj}(N - \underset{g,j}{E}X_{gj})}{N\underset{g,j}{\sigma^2}X_{gj}}\right] ,$$

which is the same as the Kuder-Richardson formula 21. Here, the subscript Q indicates the value assumed by the reliability coefficient when the only source of variability in scores is the sampling process.

Substituting (Q2) in (2) leads to the following result:

$$\text{(Q4)}\quad \rho_Q = \frac{N}{N-1}\left[\frac{\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{\underset{j}{\sigma^2}\underset{g}{E}X_{gj} + \dfrac{\underset{g,j}{E}X_{gj}(N - \underset{g,j}{E}X_{gj})}{N-1}}\right]$$

Chance Success Model: A Coefficient of Precision

Another model attributes variability in scores to chance success on multiple-choice tests. The selection of items is not regarded as a source of variability, and the results, therefore, are specific to a given test form. Let C_j be the number of items on a multiple-choice test known by person j, and let p be the probability of chance success on each of the $N - C_j$ items not known. Over repeated measurements with that particular set of items X_{gj} is binomially distributed with $\underset{g}{E}X_{gj} = C_j + p(N - C_j)$ and $\underset{g}{\sigma^2}X_{gj} = p(1-p)(N - C_j)$. Solving the first equation for C_j and substituting the result in the second gives

$$\text{(S1)}\quad \underset{g}{\sigma^2}X_{gj} = p(N - \underset{g}{E}X_{gj})\,.$$

Finding the arithmetic mean of (S1) over persons and substituting in (1), we obtain

$$\text{(S2)}\quad \underset{g,j}{\sigma^2}X_{gj} = p(N - \underset{g,j}{E}X_{gj}) + \underset{j}{\sigma^2}\underset{g}{E}X_{gj}\,.$$

Substituting the arithmetic mean of (S1) in (2) yields

$$\text{(S3)}\quad \rho_S = 1 - \frac{p(N - \underset{g,j}{E}X_{gj})}{\underset{g,j}{\sigma^2}X_{gj}}\,.$$

Here, the subscript S indicates the value of the reliability coefficient when the only source of variability in scores is chance success.

Also, substituting (S2) in (2), we obtain

$$\text{(S4)}\quad \rho_S = \frac{\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{\underset{j}{\sigma^2}\underset{g}{E}X_{gj} + p(N - \underset{g,j}{E}X_{gj})}\,.$$

From substitutions it also follows that

$$\text{(S5)}\quad \underset{g,j}{\sigma^2}X_{gj} = (1-p)^2\underset{j}{\sigma^2}C_j + p(1-p)(N - \underset{j}{E}C_j)\,,$$

that

$$\text{(S6)}\quad \rho_S = \frac{\underset{j}{\sigma^2}C_j}{\underset{j}{\sigma^2}C_j + p(1-p)(N - \underset{j}{E}C_j)}\,,$$

and that

$$\text{(S7)}\quad \rho_S = \frac{(1-p)^2\underset{j}{\sigma^2}C_j}{\underset{g,j}{\sigma^2}X_{gj}}\,.$$

Equations (S5) and (S6) were given by Carroll (2) and Horst (4). Equation (S7) indicates that, while ρ_S is equal to the variance ratio $\dfrac{\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{\underset{g,j}{\sigma^2}X_{gj}}$, it is not equal to the variance ratio $\dfrac{\underset{j}{\sigma^2}C_j}{\underset{g,j}{\sigma^2}X_{gj}}$. And, while the correlation between $\underset{g}{E}X_{gj}$ and $X_{gj} - \underset{g}{E}X_{gj}$ over all persons is zero, the correlation between C_j and $X_{gj} - C_j$ over all persons is $-\dfrac{p\underset{j}{\sigma}C_j}{\underset{g,j}{\sigma^2}(X_{gj} - C_j)}$. For other equations related to (S5), (S6) and (S7) see Carroll (2), Horst (4), and Burkheimer, Zimmerman and Williams (1).

Combined Model

Consider, now, the distribution of observed scores over repeated measurements when variability results from both of the above processes. Assume that there is, first, sampling of N items from a population, resulting in an X_{gj} value and, second, chance success on some of the $N - X_{gj}$ items in the sample not known.

Let the subscript h refer to repeated measurements where the second kind of variability is considered; then X_{ghj} is the observed score resulting from both sampling and chance success.

Over repeated measurements the values of X_{gj} will be distributed binomially as a result of sampling. For each possible value of X_{gj} the values of X_{ghj}

will also be distributed binomially, as a result of chance success, such that $\underset{h}{E}X_{ghj} = X_{gj} + p(N-X_{gj})$ and $\underset{h}{\sigma^2}X_{ghj} = p(1-p)(N - X_{gj})$. The expectation of the entire distribution of X_{ghj} values for person j is $\underset{g,h}{E}X_{ghj} = \underset{g}{E}X_{gj} + p(N - \underset{g}{E}X_{gj})$, and the variance of that distribution is

$$\underset{g,h}{\sigma^2}X_{ghj} = \underset{gh}{E\sigma^2}X_{ghj} + \underset{g}{\sigma^2}\underset{h}{E}X_{ghj}$$

$$= \underset{g}{E}p(1-p)(N - X_{gj}) + \underset{g}{\sigma^2}\left[X_{gj} + p(N - X_{gj})\right].$$

Since $\underset{g}{\sigma^2}\left[X_{gj} + p(N - X_{gj})\right] = \underset{g}{\sigma^2}\left[(1-p)X_{gj} + Np\right]$

$$= (1-p)^2\underset{g}{\sigma^2}X_{gj},$$

$$\underset{g,h}{\sigma^2}X_{ghj} = p(1-p)(N - \underset{g}{E}X_{gj}) + (1-p)^2\underset{g}{\sigma^2}X_{gj}$$

$$= p(1-p)(N-\underset{g}{E}X_{gj}) + \frac{(1-p)^2\underset{g}{E}X_{gj}(N-\underset{g}{E}X_{gj})}{N}$$

$$= \frac{(1-p)(N-\underset{g}{E}X_{gj})\left[Np + (1-p)\underset{g}{E}X_{gj}\right]}{N}$$

Also, $\underset{g}{E}X_{gj} = \dfrac{\underset{g,h}{E}X_{ghj} - Np}{1-p}$. Substituting this expression for $\underset{g}{E}X_{gj}$ in the above expression for $\underset{g,h}{\sigma^2}X_{ghj}$ gives

$$\text{(R1)}\quad \underset{g,h}{\sigma^2}X_{ghj} = \frac{\underset{g,h}{E}X_{ghj}(N - \underset{g,h}{E}X_{ghj})}{N},$$

which has the same form as (Q1). Finding the expectation of (R1) over persons and substituting in (1) also gives

$$\text{(R2)}\quad \underset{g,h,j}{\sigma^2}X_{ghj} = \frac{(N-1)\underset{j}{\sigma^2}\underset{g,h}{E}X_{ghj}}{N} + \frac{\underset{g,h,j}{E}X_{ghj}(N - \underset{g,h,j}{E}X_{ghj})}{N}.$$

Substituting in (2), we obtain

$$\text{(R3)}\quad \rho_R = \frac{N}{N-1}\left[1 - \frac{\underset{g,h,j}{E}X_{ghj}(N - \underset{g,h,j}{E}X_{ghj})}{N\underset{g,h,j}{\sigma^2}X_{ghj}}\right]$$

and

$$\text{(R4)}\quad \rho_R = \frac{N}{N-1}\left[\frac{\underset{j}{\sigma^2}\underset{g,h}{E}X_{ghj}}{\underset{j}{\sigma^2}\underset{g,h}{E}X_{ghj} + \dfrac{\underset{g,h,j}{E}X_{ghj}(N - \underset{g,h,j}{E}X_{ghj})}{N-1}}\right]$$

Here, the subscript R indicates the value of the reliability coefficient when variability in scores results from both sampling and chance success. Although equations (R1), (R2), (R3) and (R4) have the same form as (Q1), (Q2), (Q3) and (Q4), it should be noted that for given N and given values of P_j the quantity $\underset{g,h,j}{E}X_{ghj}$ does not have the same value as $\underset{g,j}{E}X_{gj}$, since $\underset{g,h,j}{E}X_{ghj} = \underset{g,j}{E}X_{gj} + p(N - \underset{g,j}{E}X_{gj})$, and that $\underset{g,h,j}{\sigma^2}X_{ghj}$ is not the same as $\underset{g,j}{\sigma^2}X_{gj}$.

From substitutions the following equations are obtained:

$$\text{(R5)}\quad \underset{g,h,j}{\sigma^2}X_{ghj} = \frac{(N-1)(1-p)\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{N} + \frac{(1-p)(N - \underset{g,j}{E}X_{gj})\left[Np + (1-p)\underset{g,j}{E}X_{gj}\right]}{N}$$

$$\text{(R6)}\quad \rho_R = \frac{N}{N-1}\Bigg[\frac{\underset{j}{\sigma^2}\underset{g}{E}X_{gj}}{\underset{j}{\sigma^2}\underset{g}{E}X_{gj} + \dfrac{\underset{g,j}{E}X_{gj}(N - \underset{g,j}{E}X_{gj})}{N-1} +}$$

$$\left. \overline{\frac{N}{N-1} \quad \frac{p(N - \underset{g,j}{E} X_{gj})}{1-p}} \right]$$

$$\text{(R7)} \quad \rho_R = \frac{(1-p)^2 \underset{j}{\sigma^2} \underset{g}{E} X_{gj}}{\underset{g,h,j}{\sigma^2} X_{ghj}} .$$

Therefore, while ρ_R as a function of $\underset{g,h,j}{E} X_{ghj}$ and $\underset{g,h,j}{\sigma^2} X_{ghj}$ is given by the same equation as ρ_Q (the Kuder-Richardson formula 21), ρ_R as a function of $\underset{j}{\sigma^2} \underset{g}{E} X_{gj}$ and $\underset{g,h,j}{\sigma^2} X_{ghj}$ is given by the same equation as ρ_S.

DISCUSSION

The method of constructing models outlined in this paper indicates how it is possible to define reliability in terms of independent, repeated measurements and at the same time recognize the specificity of various reliability coefficients which identify those measurements with different processes. As long as reliability is defined by equation (2) or its equivalent, it cannot be estimated from a single trial, as pointed out by Guttman (3). However, once the distribution function of X_{gj} has been specified, equations such as (Q3), (R3) and (S3) are obtained, which contain quantities which can be estimated from a single trial.

It has been known that the Kuder-Richardson formula 21 is a lower bound on the reliability coefficient. But when the only source of variability in scores is that assumed by the item sampling model, that formula is equal to the reliability coefficient. Similarly, when chance success is the only source of variability, the right-hand side of (S3) is equal to the reliability coefficient. The KR 21 remains a lower bound when the distribution function of X_{gj} is not known or can be only approximated.

Except for Lord's (5) derivation, the KR 21 has been obtained by first deriving the KR 20 and then assuming that all test items have the same difficulty. This assumption is implicit in Lord's model and in the procedure used in this paper, in that all items are regarded as being sampled from a common population of items. It follows that over repeated samples the expected value of an item score for person j is P_j. Over all persons and all samples the expected value of an item score is $\underset{j}{E} P_j$, and that quantity is the same for all items.

FOOTNOTE

1. The author is grateful for support received from the National Research Council of Canada during the period in which this paper was written (grant APA-252-2057-B).

REFERENCES

1. Burkheimer, G. J., Zimmerman, D. W. and Williams, R. H., "The Maximum Reliability of of a Multiple-choice Test as a Function of Number of Items, Number of Choices, and Group Heterogeneity," Journal of Experimental Education, 35:89-94, 1967.

2. Carroll, J. B., "The Effect of Difficulty and Chance Success on Correlations Between Items or Between Tests," Psychometrika, 10:1-19, 1945.

3. Guttman, L., "A Basis for Analyzing Test-retest Reliability," Psychometrika, 10:255-282, 1945.

4. Horst, P., Psychological Measurement and Prediction, Wadsworth, Belmont, California, 1966.

5. Lord, F. M., "Sampling Fluctuations Resulting from the Sampling of Test Items," Psychometrika, 20:1-22, 1955.

6. Lord, F. M. and Novick, M. R., Statistical Theories of Mental Test Scores, Addison-Wesley, Reading, Massachusetts, 1968.

7. Novick, M. R., "The Axioms and Principal Results of Classical Test Theory," Journal of Mathematical Psychology, 3:1-18, 1966.

8. Zimmerman, D. W., Williams, R. H., and Burkheimer, G. J., "Dependence of Test Reliability upon Heterogeneity of Individual and Group Score Distributions," Educational and Psychological Measurement, 28:41-46, 1968.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

READING COMPREHENSION DIFFICULTY AS A FUNCTION OF CONTENT AREA AND LINGUISTIC COMPLEXITY[1]

MILAGROS AQUINO, LUDWIG MOSBERG, and MARGE SHARRON

Southwest Regional Laboratory for Educational Research and Development

ABSTRACT

Effects of proportion of Latin suffixes and of Content Area upon newspaper article comprehension of eighth graders, as measured by cloze procedure, were investigated using a 3 x 3 factorial design. Statistically reliable mean differences were obtained between cloze test performance on Science and T. V. --Theatre passages and between such performance at High versus Medium and Low levels of Latin suffix density. The difference in mean performance across Content Areas is explained in terms of other linguistic factors--particularly differential density of prepositions.

READING COMPREHENSION may be viewed as a function of the reader's language competence, the subject matter of the language message, and the syntactic and morphological complexity of the language message itself. It has been generally recognized (6) that comprehension is differentially affected by material in different Content Areas. In terms of the above conception of comprehension this effect may be a function of the necessary background information required of the S, the author's writing style, the vocabulary specific to certain Content Areas, the interests and attitudes of the reader or listener concerning the various subject matters of Content Areas, and the linguistic and morphological complexity with which passages in different Content Areas are generally generated (3, 5, 2).

The purpose of this study was to determine whether passages concerned with different subject matter (Content Areas) differed in some dimensions of linguistic complexity and if so whether controlling these variables would result in differences in comprehension as measured by cloze test performance. The study also investigated the relationship of Ss' general reading ability and their comprehension of different Content Areas.

PRELIMINARY ANALYSIS

Five randomly chosen newspaper articles in each of ten different Content Areas from the Los Angeles imes were analyzed, using three syntactic and mor- ological variables (dependent clauses, nominal difiers, and the proportion of suffixes of Latin in). These variables were chosen because they previously (1) been found to be the best predic- s of cloze performance of fifty-nine such variables investigated. The mean frequency of occurance for each of the three variables for each Content Area is presented in Table 1. To test the reliability of the

TABLE 1

MEAN FREQUENCY OF OCCURRENCE OF THREE LINGUISTIC VARIABLES IN PASSAGES FROM TEN CONTENT AREAS

Content Area	Latin Suffixes	Dependent Clauses	Nominal Modifiers
Sports	15.0	27.2	2.2
Human Interest	15.6	31.0	5.0
Fashion	18.5	32.0	3.8
Arts	21.6	22.4	3.4
T. V. --Movies--Theatre	22.8	30.4	4.0
Religion	26.3	25.2	3.6
Government	27.2	34.4	5.6
Labor and Business	27.6	43.6	2.8
War	27.8	33.8	4.6
Science	30.4	33.0	4.8

differences in means, Duncan Multiple Range Tests were applied. The results of this analysis indicated that the proportion of suffixes of Latin origin (L-suffix density) yielded fifteen reliable differences between Content Area pairs while dependent clauses and nominal modifiers yielded five and ten reliable differences between Content Area pairs, respectively. Since L-suffix density was the most successful variable in discriminating Content Areas this variable was chosen as the independent variable for the main study. Three Content Areas were chosen for the main study on the basis that the mean of each was significantly different from the other two. A second criterion for Content Area selection was that a wide sampling of passages with varying degrees of L-suffix density could be obtained to permit a factorial study of L-suffix density effects. The three Content Areas selected were Science, Human Interest, and T. V. --Movies--Theatre.

METHOD

Materials

One hundred newspaper articles from each of the three Content Areas were taken from five daily newspapers (Los Angeles Times, New York Times, Chicago Tribune, Philadelphia Inquirer, and the Washington Post). These newspapers were chosen because they represented a wide cross section of the largest dailies and were felt to be representative of newspaper writing. Twenty articles were chosen from each newspaper in each Content Area. All articles were chosen from issues of November and December, 1967. Only the first 200 words (or the end of the sentence closest to 200 words) were considered for this study. The number of L-suffixes were counted and the proportion of L-suffixes per 200 words was calculated for each article. The criterion for determining a L-suffix was whether it was so designated by An Etymological Dictionary of the English Language (9). Proportions ranged from .000 to .200. Frequency distributions for each Content Area indicated that proportions tended to cluster at the high end for Science and at the low end for T. V.--Movies--Theatre with Human Interest in the middle. On the basis of the distributions, proportions of .01 to .05 were classified as "Low," .06 to .10 as "Medium," and .11 to .15 as "High." Ten articles from each Content Area at each level were chosen at random with the restriction that two articles would be chosen from each of the five newspapers for each Content Area and level. It was not possible to meet this criterion exactly but only a few exceptions occurred. In some cases it was not possible to obtain two articles at either the "High" or "Low" level in a content area from a given newspaper.

PROCEDURE

The articles chosen for the final study were made into cloze tests by deleting every fifth word and replacing it with a blank of standard length. Each article was tested on five Ss. Each S receiving a passage was given a cloze test wherein deletions began with either the first, second, third, fourth, or fifth word. Therefore, every word of each article was tested once. Each S read and responded to two articles during a 1-hour testing period. Articles were randomly paired with the restriction that they both be in the same Content Area and at the same level of L-suffix density. Ss were tested in groups of approximately 25. E read the instructions to the class and negotiated a sample test with the Ss. All Ss were tested by one E and were given 40 minutes to complete the two passages. Stanford Reading Achievement Test scores on all Ss were obtained from school records. These tests were administered approximately 6 months prior to this study. Twenty-five Ss were tested under each of the nine treatment conditions investigated.

SUBJECTS

The Ss were 225 eighth graders from one Southern California school district. The district is a predominantly lower-middle to upper-middle class community. The Stanford Reading Achievement Test (SRAT) scores indicated that the sample included a wide range of reading ability.

RESULTS

Cloze items were scored correct only if an insertion was the exact word deleted; accurate spelling was not required. The total score for each S was calculated by combining the two passages negotiated by each S and calculating the ratio or correct insertions over the total number of deletions.

To ensure that the random assignment of Ss to Content Area and levels of L-suffix density did not result in group differences in reading ability, the SRAT scores were compared. A 3 x 3 analysis of variance showed no reliable differences between groups. The means are presented in Table 2. The means and the analysis of variance give strong indication that the experimental groups were comparable on SRAT scores.

TABLE 2

MEAN SRAT SCORES AT EACH CONTENT AREA AT EACH LEVEL OF LATIN-SUFFIX DENSITY

Latin Suffix Density	Content Area			
	Science	Human Interest	T. V. --Movies--Theatre	Mean
High	8.20	8.34	8.58	8.37
Medium	9.10	8.07	7.80	8.39
Low	8.89	8.20	8.26	8.
Mean	8.73	8.20	8.28	

TABLE 3

MEAN CLOZE PERFORMANCE AS A FUNCTION OF CONTENT AREA AND LEVEL OF LATIN-SUFFIX DENSITY (MEAN PROPORTION CORRECT)

Level of Latin-Suffix Density	Content Area			
	Science	Human Interest	T. V. --Movies--Theatre	Mean
High	.290	.292	.280	.287
Medium	.346	.374	.292	.337
Low	.398	.303	.302	.335
Mean	.345	.323	.292	

Mean cloze test performance data are presented in Table 3. The mean proportion of correct responses is highest for Science and lowest for T.V. --Movies--Theatre. Mean performance was the same for Low and Medium levels of L-suffix density and somewhat lower for the High level. A 3 x 3 analysis of variance (Table 4) indicated significant differences for both main effects (Content Area, L-suffix density). Specific comparison of means using the Duncan Multiple Range Test showed that the only reliable difference between Content Areas was between Science and T.V.--Movies--Theatre. For L-suffix density, the High mean was significantly lower than either the Low or Medium means.

TABLE 4

ANALYSIS OF VARIANCE OF CLOZE SCORES

Source	df	SS	MS	F
Content Area(CA)	2	.107	.054	3.176*
Latin-Suffix Density (LS)	2	.118	.059	3.471*
CA x LS	4	.131	.033	1.941
Error	216	2.509	.017	

* $p < .05$

The correlations between SRAT and cloze scores are presented in Table 5. Separate correlations were computed for each Content Area collapsed over L-suffix density levels and for each level of L-suffix density collapsed over Content Area. Comparable correlations (.612 and .644) were obtained for Human Interest and T.V. --Movies--Theatre but a somewhat lower correlation between Science and SRAT (.515). Similarly, comparable correlations were obtained between SRAT and Medium and Low L-suffix density levels (.651 and .610) but a somewhat lower correlation between High and SRAT (.508).

To shed further light on the difference in cloze performance between Science and T. V. --Movies--Theatre, several ad hoc analyses of the passages were performed. An analysis of the mean number of sentences per passage and the mean number of words per sentence revealed that Science passages contained fewer but longer sentences than T. V. --Movies--Theatre. The mean number of sentences per passage was 8.86 and 10.90 for Science and T.V.--Movies--Theatre, respectively. The mean number of words per sentence was 22.98 and 18.86 for the two Content Areas respectively. Analysis of variance indicated that both of these differences were statistically significant at the .05 level [$F(2,87)=5.39$], [$F(2,87)=3.42$]. Certain modifiers such as prepositional phrases and relative clauses make sentences longer. It was hypothesized that since Science passages had longer but fewer sentences they would contain more prepositions. The mean number of prepositions in Science passages was 29.0 and 23.7 for T. V. --Movies--Theatre. Analysis of variance indicated that this difference was reliable at the .01 level [$F(1,58)=23.40$]. Prepositions are essentially relater words which signal a relationship of dependency for the object of the preposition with respect to the item modified. As such, prepositions are often uniquely determined by the context. Therefore, Science passages having more prepositions, are easier for the Ss when comprehension is tested by the cloze procedure. Further inspection of the passages of these two Content Areas suggested that T.V.--Movies--Theatre passages were generally written in more colloquial language and contained more proper nouns and compounds.

DISCUSSION

The result that there is a reliable difference in cloze performance between the High level of L-suffix density and both Medium and Low levels suggests that passages in content areas which normally contain a high proportion of L-suffixes should be more difficult in terms of cloze performance than passages in content areas in which few L-suffixes normally appear. Preliminary analyses of Content Areas indicated that Science passages from newspapers contain significantly more L-suffixes than either of the other two Content Areas under investigation. The fact that Science passages give the highest cloze performance when L-suffixes are controlled suggests that the proportion of L-suffixes is a reasonably powerful predictor of cloze difficulty. However, the fact that cloze performance was not differentially affected by the Low and Medium levels suggests that a rather high level of L-suffixes is necessary before this variable appreciably affects cloze scores. It must also be pointed out that as a variable, L-suffix density covaries with several other variables. Words containing L-suffixes tend to be longer words than non-L-suffixed words and typically less frequent or familiar in the language.

TABLE 5

CORRELATIONS BETWEEN CLOZE SCORES AND SRAT SCORES AT EACH LEVEL OF LATIN-SUFFIX DENSITY AND AT EACH CONTENT AREA

Treatment	Correlation
Science	.515
Human Interest	.612
T. V. --Movies--Theatre	.644
High	.508
Medium	.651
Low	.610

The reliable superiority of cloze performance on Science to that of T.V.--Movies--Theatre seems to be a function of several variables. Since Science articles contained significantly longer and fewer sentences but an equal number of words, there were a greater number of prepositions in the Science passages. In terms of cloze tests, function words are generally easier than content words because the former, especially prepositions, are uniquely determined by the linguistic context. Other factors contributing to the observed Content Areas difference were that the T.V.--Movies--Theatre passages were generally more colloquial; contained more proper nouns and compounds and were often concerned with more than one topic per passage. These factors would have the effect of reducing context clues and both syntactic and semantic redundancies.

Results of the correlational analyses between SRAT and cloze scores by Content Areas and L-suffix density levels indicated that the correlations are well within the range of previous investigations (4, 7, 8). Although a somewhat lower correlation was obtained for Science than the other two Content Areas, reading ability does not appear to be differentially related to differences in cloze performance over Content Areas.

FOOTNOTE

1. This research was supported pursuant to a contract with the United States Department of Health, Education and Welfare, Office of Education.

REFERENCES

1. Aquino, Milagros, "Development of Linguistically-based Readability Formula," Southwest Regional Laboratory Research Memorandum, 1968.

2. DeVito, J., "Comprehension Factors in Oral and Written Discourse of Skilled Communicators," Speech Monographs, 32: 124-128, 1965.

3. Farr, J.N., Jenkins, J. J. and Paterson, G., "Simplification of Flesch Reading Ease Formula," Journal of Applied Psychology, 35:333-337, 1951.

4. Fletcher, J. E., "A Study of the Relationships Between Ability to Use Context as an Aid in Reading and Other Verbal Abilities," unpublished doctoral dissertation, University of Washington, 1959.

5. Gunning, R., The Technique of Clear Writing, McGraw-Hill, New York, 1952.

6. King, W. H., "An Experimental Investigation into the Relative Merits of Listening and Reading Comprehension for Boys and Girls of Primary School Age," Journal of Educational Psychology, 29: 42-49, 1959.

7. Mosberg, L., Potter, T. C. and Cornell, R. K., "The Relation Between Choice and Multiple-choice Test Scores as a Function of Relative Paragraph Difficulty and Grade Level," Southwest Regional Laboratory, Technical Report No. 6, 1968.

8. Ruddell, R.B., "The Effect of Oral and Written Patterns of Language Structure on Reading Comprehension," Reading Teacher, 18:270, 1965.

9. Skeat, W. W., An Etymological Dictionary the English Language, Oxford University Pre London, 1963.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

EFFECTS ON STUDENT ACHIEVEMENT OF BEHAVIORAL AND NONBEHAVIORAL OBJECTIVES[1]

EVA L. BAKER

University of California, Los Angeles

ABSTRACT

Educators disagree on the relative merits of stating classroom objectives behaviorally or nonbehaviorally and have done little to add data to their argument. An experiment was conducted in the field of social science where one of three lists of objectives--one list nonbehavioral, the other two behavioral--was randomly assigned to participating high school social studies teachers who were instructed to teach objectives in their classes. Unit sampling was used and eighteen classrooms were involved. Students were measured, using a form of item sampling, on the acquisition of the five skills stated in the behavioral objectives as well as on eighteen transfer skills. Teachers' faulty understanding of objectives, indicated by their inability to provide relevant classroom practice and to identify, when asked, test items measuring given objectives, may have accounted for lack of differences.

THERE IS widespread interest in the educational world regarding how one might best state objectives for instructional effectiveness. The controversy centers about the use of objectives stated in operational terms, specified with respect to the behaviors the student will be able to demonstrate after instruction. Contestants have argued with varying degrees of eloquence either favoring or opposing the precise explication of objectives. Proponents cite the obvious value in the teacher's knowing where he is going and how to evaluate when he gets there (4, 6, 9, 11). Others (1, 2, 3, 8) worry that precise objectives hinder the full development of the student and force the teacher into an inflexible mode. The argument is ceasing to be academic with the advent of the Program Planning and Budgeting System (PPBS) in school finance and the dependence of such a system on operationally defined and measureable effects of instructionally related expenditures.

Unfortunately, there has been little concrete research conducted into the effects on students when teachers use behavioral or nonbehavioral objectives. McNeil (10) investigated the effects achieved when student teachers were told that their grade would be determined by the effects they produced in their learners rather than the usual means-referenced observation scales. His results favored teachers in the ends-oriented rather than means-oriented groups, but because the results were related primarily to the student teachers' motivation for a grade, the findings e not directly applicable to the behavioral objectives ntroversy.

Perhaps the major difficulty with empirical investigation has been in conceiving of an adequate design for an adequate experiment. Because most teachers have disparate objectives, comparisons between the effects they produce in their students are difficult. Problems are encountered if one randomly provides one group of teachers with a set of behavioral objectives and another group with nonbehavioral objectives and observes the effects on the students' performance of the behavioral objectives. Such a model is clearly biased against the teachers who received nonbehavioral objectives since they might have thought of observable ways to test the nonbehavioral objectives other than those hit upon by the researcher.

The research reported herein sought to eliminate this design problem by allowing the comparison between the effects of teachers' use of behavioral and nonbehavioral objectives to be made on criterion measures which took account of interpretations of the nonbehavioral objective other than the specific behavioral objectives used in the investigation.

SUBJECT AND SAMPLING

Subjects were eighteen experienced teachers of social studies in Nevada secondary schools. These teachers were urged to participate in the evaluation of a new instructional unit by their principals. Teachers were assigned at random to one of three treatment groups.

MATERIALS

The instructional materials used in the experiment consisted of a resource unit in the field of social science research methodology. The topic of the unit was chosen because it could fit in anywhere in the curriculum and would be relatively free of the contamination of previous instruction, since the topic is not typically in the curriculum. The unit consisted of a list of five objectives, expository material for the teachers' use to refresh themselves in the subject matter, possible activities, and abstracts of actual investigations conducted in the broad field of social science. The activities were constructed so that some were clearly relevant to the behavioral objectives and some were irrelevant to those objectives.

TREATMENT

The design of this study attempted to deal with the problem of multiple interpretations of nonbehavioral objectives by teachers. Instead of comparing student performance only on the particular behavioral objectives given to one group of teachers, the comparison was made also on the basis of other behavioral interpretations of the nonbehavioral objectives. These other interpretations were made by four graduate students in social science. For example, one nonbehavioral objective was: To Understand The Difference Between Descriptive and Causal Hypotheses. Four different behavioral objectives dealing with such "understanding" were constructed.

Two examples were:

1. When given a hypothesis the student can identify it as either descriptive or causal.

2. Student will be able to write a sentence contrasting the use of descriptive and causal hypotheses.

Teachers in both "behavioral objective" treatments received only one of the behavioral sub-objectives to teach, but students in all groups were measured on their achievement of all such objectives. Such a procedure takes into account at least some of the different interpretations of a given nonbehavioral objective.

In this study, five nonbehavioral objectives on five distinct topics in social science research methods were written. For each of the nonbehavioral objectives several different behaviors were described which might be reasonable interpretations of the nonbehavioral objectives. Two of the nonbehavioral objectives had four behavioral interpretations, and the other three nonbehavioral had five behaviors each, resulting in a list of twenty-three behavioral objectives.

Given the twenty-three behavioral objectives generated from the original five nonbehavioral objectives, a representative list of behavioral objectives had to be selected. Since the total number of possible combinations of five behavioral objectives which could be selected from the list of twenty three behavioral objectives was uncomfortably large, one way to secure a representative list of five behavioral objectives was to select randomly from within each topic one objective to be used. As insurance against the possibility that this single random collection of behavioral objectives was somehow typical of peculiar by chance, another set of five behavioral objectives was formed by asking instructional experts to select within each of the five categories the one behavioral objective they felt would facilitate learning of all other objectives within the same category. Experts were asked to rank each objective within each topic area and in each case the modal rank was selected.

Each treatment, then consisted of a separate list of five objectives. Each objective matched with objectives on the other lists in terms of content (objective one on all three lists, for example, dealt with hypotheses); however, one list was composed of behavioral objectives selected at random (RBO) the second list was composed of behavioral objectives selected by instruction experts (EBO), and the last list was composed of the nonbehavioral objectives (NBO).

INSTRUMENTATION

Test items for each of the twenty-three subobjectives were constructed. All items were item analyzed based on trials with the material with fifty-seven high school students. The Kuder Richardson (formula 21) reliability index computed was .81. Additional reliability data were obtained prior to the experiment by having a teacher in a Los Angeles area junior high school teach the behavioral objectives and administer a posttest based on the objectives. The reliability coefficient using the Kuder Richardson 21 formula was .83.

Item sampling (5 and 7), an approach to gather many observations in a relatively limited time period, was employed since a representative test measuring all twenty-three sub-objectives would require too much time to administer in a normal school setting. Item sampling in this case entailed giving different short but comparable forms of the test to learners within a given classroom. Four forms of the test were used so that the tesing period could be contained in one class period. While some objectives could be adequately measured by one item ("Describe the difference between causal and descriptive hypotheses"), other objectives require more than one item ("Given a hypothesis, label it as either causal or descriptive hypotheses"). However, the test items were randomly assigned to the four test forms within each of the twenty-three behavioral objectives, so that each test contained the same number of items measuring a given objective as any other form. The test forms were non-overlapping, except for those objectives for which only one item could sensibly be constructed. Provisions were made so that tests were randomly assigned within each classroom, which resulted in one quarter of the students in a given classroom taking any one form of the test. While different learners within a classroom received one of four versions of the posttest, the total effect was that each teacher's class received all items devised for the twenty-three behavioral objectives in the same proportion as any other class. Therefore, while individual students' performances within and between classes were not necessarily comparable, the teachers' instructional effectiveness, as gauged by the mean performance of the class, could be compared.

PROCEDURE

The eighteen teachers had been assigned at random to one of the three treatments. Each teacher received a list of five objectives, the resource unit, a sealed test packet and a questionnaire. Teachers in the nonbehavioral objectives (NBO) group received five objectives which specified the content areas clearly but did not describe the specific behaviors to be tested. Teachers in the random behavioral objectives (RBO) group received a list of five objectives, selected at random from within each of the five major content areas from a list of twenty-three specific objectives. The instructional expert behavioral objectives (EBO) group received a list of five behavioral objectives selected from the twenty-three objectives by a team of individuals reputedly expert in the field of instruction.

Teachers were told that they were helping to assess the usefulness of the objectives and resource materials with their students and that results by teacher would not be given to the district or supervisors. Teachers were instructed to teach the five objectives within two class periods (or taking about 100 minutes if their classes were on flexible scheduling basis). The tests were administered immediately following instruction. Each test packet contained the four test forms in random order. The teachers were asked to distribute the tests in that order as they removed them from the sealed packet. Teachers were asked to return to questionnaires and tests within one week of completing the unit.

ANALYSES AND RESULTS

The students' tests were scored and weighted so that each of the twenty-three objectives made an equal contribution. A comparison among all three groups on the total test score, reflecting twenty-three objectives, was conducted through use of one-way analysis of variance. Separate analyses of variance were conducted for the three groups by each of the five topic areas. Analyses were also conducted comparing the random behavioral group and the nonbehavioral group and between the expert behavioral group and the nonbehavioral group on their respective lists of objectives. Results of the analyses yielded no significant differences. The data of real interest, however, were gathered from the teacher questionnaire. Teachers were asked to select test items which directly measured the objectives they had been given. It was believed that a teacher who carefully understood the objectives should be able to match 75 percent of items with their proper objectives. In Table 1 is the average number of correct responses made by teachers in the behavioral and expert behavioral groups.

TABLE 1

MEAN AND STANDARD DEVIATION FOR SCORES FOR MATCHING 12 ITEMS WITH GIVEN BEHAVIORAL OBJECTIVES

Sample Size	13
Mean	6. 12
Standard Deviation	1. 77

As can be seen, most teachers were responding to these binary choices at approximately chance levels (51 percent). Such information is extremely important when considering the implications of the pupil achievement data.

The questionnaire also called for the teacher to record the type of activities he used and time allocated to each. Activities chosen or described were judged either appropriate to the objective or inappropriate and a percentage of appropriate activity time compared with overall time was computed for each teacher in both the random behavioral and expert behavioral groups. No such percentage was computed for the nonbehavioral teachers since, given nonbehavioral objectives, one might infer that almost any activity dealing with the content area would be appropriate to the objective. Percentages for the random behavioral and expert behavioral groups are presented in Tables 2 and 3.

TABLE 2

PERCENTAGE OF INSTRUCTION TIME JUDGED TO BE ALLOCATED TO RELEVANT ACTIVITIES BY BEHAVIORAL GROUPS

Sample Size	13
Mean	. 2795
Standard Deviation	. 3458

TABLE 3

PROPORTION OF TOTAL NUMBER OF REPORTED ACTIVITIES JUDGED TO RELEVANT TO GIVEN OBJECTIVES

Sample Size	13
Mean	. 2729
Standard Deviation	. 3112

DISCUSSION

The retrospective explanations reasonably justified by this configuration of data are straightforward. A major inference one must make is that the treatment variations as presented were not powerful enough to modify the dependent variable performance. This is reflected in the within group variability of the teachers. However, the treatment in the behavioral objective groups did not "take" as expected. For example, one would imagine that a teacher who really attended to the precise nature of the objectives (behavior and content) would be able to identify those test items which were clear measures of the objectives. The data reveal that this was not the case in both behavioral groups.

Further support for the contention that the behavioral objectives treatment had no major impact on the teachers is supplied by the data obtained from teacher activity reports, where the teacher was supposed to indicate which activities he selected during the instruction. Estimated time devoted to each activity was also reported. Proportions were compiled for time spent in activities over total time reported and for the number of activities judged relevant (by trained evaluators) over total number of activities reported. The figure for the behavioral groups on the time percentage was approximately .28. On the number of relevant activities proportion the groups averaged .27.

If one believes that behavioral objectives should make a difference in pupil learning, one implication of this finding would support a replication in which a group of teachers thoroughly trained in the use of behavioral objectives would be contrasted on the basis of their students' achievement with an untrained group of teachers. In addition to expanding the treatment time or increasing the number of participating teachers, another important factor may have to be introduced in future studies of this type. Even if teachers do understand what behavioral objectives are, one still must assess the extent to which teachers are committed to producing pupil achievement. Teachers may not be intrinsically rewarded for securing specified pupil behavior change, and such a motive would have to be strong enough to cause them to change routine, and possibly, less efficient, instructional methods. Research into the application of suitable contingency models for reinforcing measurable changes in student behavior is essential.

FOOTNOTE

1. The materials used in this research were developed in the course of Project No. 5-0566-2-12-1 supported by the Office of Education, U. S. Department of Health, Education and Welfare.

REFERENCES

1. Arnstine, Donald G., "The Language and Values of Programmed Instruction: Part 2," The Educational Forum, 28, March 1964.
2. Atkin, Myron J., "Some Evaluation Problems in a Course Content Improvement Project," Journal of Research in Science Teaching, 1:129-132, 1963.
3. Eisner, Elliot W., "Educational Objectives: Help or Hindrance?" paper presented at 50th Annual Meeting, American Educational Research Association, Chicago, Illinois, February 17-19, 1966.
4. Gagne, Robert M., "Curriculum Research and the Promotion of Learning," Perspectives of Curriculum Evaluation, A. E. R. A., 1967.
5. Husek, T. R. and Sirotnik, Ken, Item Sampling in Educational Research, Center for the Study of Evaluation of Instructional Programs, Los Angeles, California, Occasional Report No. 2, February 1967.
6. Lindvall, C. M., "The Importance of Specific Objectives in Curriculum Development," Defining Educational Objectives, A report of the Regional Commission on Educational Coordination and The Learning Research and Development Center, Pittsburgh, Pennsylvania, 1964.
7. Lord, F. M., Item Sampling in Test Theory and In Research Design, Princeton, New Jersey, ETS, RB-65-22, 1965.
8. MacDonald, James B., "Perspective on Technological Rationality in Education," paper presented at the Association for Supervision and Curriculum Development, Curriculum Research, Eastern Section, Washington, D. C., November 3-5, 1966.
9. Mager, Robert F., Preparing Objectives for Programmed Instruction, Fearon Publishers Inc., San Francisco, 1962.
10. McNeil, John D., "Concomitant of Using Behavioral Objectives in the Assessment of Teacher Effectiveness," paper presented at the 50th Annual Meeting, American Educational Research Association, Chicago, Illinois, February 17-19, 1966.
11. Tyler, Ralph W., Basic Principle of Curriculum and Instruction, Chicago, Illinois, University of Chicago Press, 1950.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

VIDEOTAPE FEEDBACK AND MICROTEACHING IN A TEACHER TRAINING MODEL[1]

WALTER R. BORG, WARREN KALLENBACH, MERVA MORRIS, and ALLEN FRIEBEL
Far West Laboratory for Educational Research and Development,
University of Utah, San Jose State College

ABSTRACT

The goals of this study were to determine the effectiveness of the minicourse model in changing student teacher behavior and to estimate the effects of the microteaching format and use of videotape feedback within this model. Feedback and practice in the microteaching format were manipulated with four groups of student teachers while a fifth group served as a control.

Eleven behaviors were scored for the four treatment groups on precourse and post-course videotapes. Eighteen of the resulting comparisons were significant beyond the .01 level and three beyond the .05 level in the desired direction. Three of the eleven control group gains were significant at the .05 level. Treatment groups that did not receive videotape feedback and did not practice in the microteaching format were not significantly different than groups that did.

THE MINICOURSE MODEL

SINCE March, 1967, the primary program of the Far West Laboratory for Educational Research and Development has been aimed at developing and testing the minicourse instructional model as a tool for changing the classroom behavior of teachers. The minicourse model was initially based upon the microteaching technique used at Stanford University in the training of teacher interns. Microteaching as developed at Stanford has several characteristics:

1. First, a set of specific teaching skills is studied by the intern.
2. Then, the intern attempts to apply the skills in a short lesson, usually 5 to 10 minutes, with four or five pupils.
3. This lesson is recorded on videotape and immediately after its completion, the intern watches a replay of the lesson.
4. During the replay a specially trained supervisor gives the intern specific feedback on his performance in the skills.
5. The intern then replans the lesson and reteaches it to another group of four or five pupils.

The minicourse model differs from the Stanford intern model in several ways which generally reflect the different situation found in a university preservice program as opposed to an inservice program that may be offered in any school district. One major difference is that the minicourse model provides a self-contained package of inservice training materials that can be used in any school where videotape recording system is available. The second major difference is that while the Stanford Intern Program employs feedback from trained supervisors, the minicourse program attempts to provide feedback through carefully structured teacher self-evaluation of televised lesson replays. The minicourse also relies heavily upon filmed illustrations by model teachers rather than supervisory feedback to provide the trainee with a basis for discriminating the behavior patterns or skills to be learned. Some research evidence raises serious doubts about the value of supervisory feedback. Acheson (1), in a study of sixty teaching interns at Stanford, found that television feedback with a supervisor conference was no more effective than television feedback without a supervisor conference. Furthermore, a supervisor conference without television feedback did not produce greater behavior changes than occurred with a control group which received neither television feedback nor conference. In a study of the effects of student versus supervisor feedback on teaching methods, Tuckman and Oliver (12) found that supervisor feedback resulted in a negative shift in teacher behavior, i.e., away from the direction suggested by the supervisor.

On the other hand, evidence on the use of television replays as a source of feedback, such as Acheson's (1) study, is generally positive. Orme's (11) study of perceptual modeling versus symbolic modeling is also pertinent to the minicourse instructional model. In symbolic modeling the subject was told what to do while in perceptual modeling, he was shown a filmed model who portrayed the desired behavior. Perceptual modeling consistently brought about greater changes in the behavior of teaching interns than symbolic modeling. Using models in place of supervisory feedback also has the practical advantage of making it possible for the minicourse to be developed as a complete self-contained package which requires no special skills to administer.

PROBLEM

The major goals of this study were to estimate the degree to which (a) practice in the microteaching format (i. e., short lessons and few pupils) and (b) feedback from the videotape replay influenced learning in the minicourse model. A third goal was to determine the effectiveness of the minicourse as a technique for changing the behavior of student teachers.

The specific hypotheses that were tested in this study are:

1. Student teachers completing the entire Minicourse 1 will display significantly greater changes in the teaching behaviors covered in Minicourse 1 than a similar group which receives the entire course except for practice in the microteaching format and VTR feedback.
2. Student teachers completing the entire Minicourse 1 sequence will display significantly greater changes in the teacher behavior covered in Minicourse 1 than student teachers who complete Minicourse 1 without VTR feedback.
3. Student teachers taking the minicourse in conjunction with student teaching will make significantly greater changes in the teaching behaviors covered in Minicourse 1 than similar subjects taking student teaching without the minicourse.

The microteaching format appears to have some advantages over practice in the whole class situation. Bush and Allen (6) suggest that microteaching permits the teacher to practice new skills and try new ideas in a less difficult situation than the regular classroom. This reduces the likelihood of failure and the threat to the teacher that is implicit in trying new approaches. Research evidence indicates that skills learned in the microteaching format transfer to a significant degree to the teacher's behavior in his regular classroom (4) and persist with little or no regression for a period of several months (5). There is also some evidence (10, 3) to indicate that microteaching achieves changes in teacher behavior much more rapidly than student teaching or intern teaching.

There is some theoretical basis for expecting that the role of videotape feedback might be a major factor in the success of the minicourse model. This feedback is focused upon the specific skills that the teacher is trying to master. The videotape replay is viewed by the teacher within a few minutes after he completes the lesson, thus providing him with almost immediate information on his performance. However, although the videotape replay appears to be an important factor in the minicourse model, there is no direct evidence on the amount it contributes to the behavioral changes brought about by this model. Since many schools do not have access to a videotape recording system, it is important to learn whether the minicourse model can bring about significant behavior changes when used without the videotape recorder (VTR).

PROCEDURE

Developing Minicourse 1

The development cycle employed at the Laboratory calls for three field tests of each minicourse, each followed by a revision based on field test information. The preliminary inservice form of Minicourse 1 was developed in the spring of 1967 and field tested in July of that year. In the Laboratory's R and D cycle, the purpose of the preliminary field test is to get qualitative feedback from Laboratory personnel who conduct the test and from participating teachers. Based on the results of this field test, the course was revised and a main field test was conducted in October. This test was designed to determine if the minicourse brought about significant changes in the skills and behaviors covered in the course. The performance of the forty-eight participating teachers was videotaped in 20-minute pre-course and post-course lessons recorded under essentially the same conditions. The tapes were then assigned randomly to trained raters who replayed the tapes and scored them on the behaviors taught in the course. Raters were not regular Laboratory employees, had no stake in the success or failure of the program, and did not know whether any given tape was made before or after the course. Main field test results on the inservice teacher sample were reported in a previous paper (4). Briefly, it was found that of the thirteen scores obtained from the pre-course and post-course tapes, eleven showed statistically significant gains. Some of the gains were not only statistically significant, but were large enough to suggest a major change in the way the participating teachers conducted their discussion lessons. For example, the average percentage of teacher talk during class discussion changed from 51.64 percent on the pre-course tapes to 27.75 percent on the post-course tapes.

Minor revisions were made after the main field test and a final inservice field test was conducted February, 1968, to determine whether the course could be conducted by school personnel with a minimum of Laboratory help, i. e., was it ready for operational use? The operational field test was successful, and only minor revisions were required before the course was made available to the schools for operational use in April, 1968.

The operational test form of Minicourse 1 was adapted for preservice use in the study reported in this paper.

THE TREATMENT

Minicourse 1 was initially developed as an inservice course for teachers in grades 4, 5, and 6. The goal of the course is to change twelve specific classroom behaviors related to the teacher's method of conducting a discussion lesson. These twelve behaviors are organized into four instructional sequences. The objectives of each instructional sequence and the three behaviors covered in each are given in Table 1. An instructional sequence in Minicourse 1 requires three 75-minute periods to complete. During the first period, the student teacher subgroups that were given the entire course started by watching the instructional lesson, a 20- to 30-minute motion picture. In this film three behaviors are described by a narrator and illustrated with film clips which show teachers using the behaviors. A rationale for using each behavior is also given although this is treated in more depth in a handbook given the student teacher at the start of the course. Immediately after watching the instructional film, the student teacher watched a model film in which a model teacher illustrates all three of the behaviors in a microteaching situation, i. e., a lesson 5 to 10 minutes long involving four to eight pupils. On the first showing of the model lesson, the student teacher was asked to identify each of the three behaviors on a checklist as it occurred. A second form of the model lesson was then shown in which the narrator named each behavior as it occurred so that the student teacher could check his ability to recognize the behaviors. After viewing the model film the student teacher was given instructions for preparing a short discussion lesson in which he could practice the three behaviors described in the instructional and model films.

During the second period in the instructional sequence the student teacher took four to eight pupils from his assigned class to the microteaching room, started the videotape recorder and taught the lesson he had planned. When he had finished teaching, he sent the pupils back to the regular classroom and replayed twice the videotape recording of his lesson. For each of these replays, an observation form was provided which focuses the student teacher's attention on the three specific behaviors he was instructed to practice. After viewing the replays of his lesson the student teacher was instructed to replan the same lesson so as to make more effective use of the behaviors practiced.

At the start of the third session, the student teacher again went to the microteaching room and taught his revised lesson to a different group of pupils from his assigned class. The lesson was again recorded on videotape which the student teacher replayed twice, each time using a different observation form that he focuses on a specific teaching behavior.

Having completed the first instructional sequence, he was then ready to proceed to the next sequence in which he followed the same pattern of viewing instructional and model films, microteaching and replaying a lesson, replanning the lesson, and reteaching and replaying the revised lesson.

TABLE 1

OBJECTIVES AND BEHAVIORS FOR EACH INSTRUCTIONAL SEQUENCE IN MINICOURSE 1

Instructional Sequence I:

Objective: To change teacher behaviors that will increase the pupil's readiness to respond to discussion questions.

Specific behaviors covered:
A. Ask question, pause 5 seconds, then call on pupil.
B. Deal with incorrect answers in an accepting, nonpunitive manner.
C. Call on both volunteers and non-volunteers.

Instructional Sequence II:

Objective: To change teacher behavior so as to decrease teacher participation and raise the level and amount of pupil participation.

Specific behaviors covered:
A. Redirection-directing the same question to several pupils.
B. Framing questions that call for longer pupil responses and preclude one word replies.
C. Framing questions that require the pupil to use higher cognitive processes.

Instructional Sequence III:

Objective: To increase the teacher's use of probing behaviors in order to guide the pupil to more complete and thoughtful responses.

Specific behaviors covered:
A. Prompting.
B. Seeking further clarification and pupil insight.
C. Refocusing the pupil's response.

Instructional Sequence IV:

Objective: To reduce the frequency of teacher behaviors that interfere with the flow of the discussion.

Specific behaviors covered:
A. Teacher should not repeat her questions.
B. Teacher should not answer her own questions.
C. Teacher should not repeat pupil answers.

SUBJECTS

In this study, student teachers from three teacher training institutions were employed as subjects. In College A, seventeen student teachers completed the entire Minicourse 1 as described in the previous section (Group A-1). A second group of sixteen student teachers (Group A-2) were given the course handbook, shown the instructional and model films, and were instructed to practice the behaviors in their regular student teaching classrooms. This group did not practice in a microteaching situation and did not view replays of their performance.

In College B, fifteen student teachers completed the entire Minicourse 1 (Group B-1). A group of seventeen student teachers were given the entire course except for the videotape recording and replay (Group B-2). This group received the handbook, was shown the instructional and model lesson, and practiced in a microteaching situation. They also evaluated their microteaching lessons but since they did not view themselves on videotape, the evaluation forms developed for this group were less specific.

In College C, a group of student teachers taking Minicourse 1 were to be compared with a control group which received regular student teaching only. However, a sufficient number of student teachers could not be assigned to the two treatments so it was decided to collect pre- and post-course videotapes on the control group only, and employ this as a comparison group for the other two colleges (Group C-1).

In Colleges A and B student teachers scheduled to teach in grades 4, 5, and 6 could not be assigned to the treatments in a completely random fashion. In College A, those student teachers participated who had been enrolled with four college supervisors. Student teachers were assigned arbitrarily to supervisors. Which treatment the student teacher received was determined by the school to which he was assigned for student teaching--another arbitrary administrative decision.

College B student teachers had received their student teaching assignments before arrangements were made to participate in the study. Student teachers in schools which had four student teachers assigned to grades 4, 5 or 6 were placed in Group B-1. Student teachers in schools which had two or three assigned to grades 4, 5 or 6 were placed in Group B-2. This decision was made to get maximum use out of the video equipment, since up to four teachers in a given school can use the same set of equipment. No variables that would bias the assignment of student teachers to the treatments appeared to be operating in either college. In College C, all student teachers assigned to five elementary schools were assigned to the comparison group. Again, no apparent bias was operating in this selection.

The dependent variable in this study was the subject's performance on two 15-minute videotapes recorded in a regular classroom with the student teacher's entire class; one immediately before and one immediately after the experimental treatment. A week before the pre-course tape was recorded and again before the post-course tape was recorded, subjects were given instructions on how to prepare their discussion lessons for videotape recording. During the recording the camera was placed in the classroom in a position that would provide a good view of the student teachers and pick up at least half of the pupils. Four microphones were positioned around the classroom. At College A the videotape recorder, monitor and operator were positioned outside of the classroom, while in College B they were positioned at the rear of the room. The student teachers were given a brief warm-up period before the recording was started.

There seems little doubt that the presence of the camera and microphones changes the classroom situation. However, there is some evidence that this change is small after pupils have had a few minutes to adjust to the presence of the equipment (9). In this study an effort was made to minimize this influence and hold it constant for all subjects in a given college. Other alternatives, such as using a hidden camera or leaving the camera in the room for several days and recording on a time sampling basis, were considered, but could not be used because of limitations in the equipment and classrooms that were available.

After the experimental treatment, the pre-course and post-course tapes were coded so the raters would not know when a given tape had been recorded. Raters were trained to reliably score two or three specific behaviors at a time. During the training sessions, a very precise operational definition of each behavior was developed and decisions were made on the scoring of borderline cases. The fifteen raters were then assigned randomly to three groups. Each group was assigned five tapes, and raters in the group independently scored these tapes. The average inter-rater reliability was then computed. Table 2 gives the average correlations among the five raters in each group.

The factor that appeared to have the greatest influence upon the reliability coefficient was the frequency of occurrence of the behavior. Reliability coefficients for behaviors that occurred with low frequency such as answering own questions, prompting, and asking higher cognitive questions, were lower and less stable from group to group than for high frequency behaviors such as redirection and repeating pupil answers. The specificity of the behavior also influenced reliability since it is much easier for an observer to determine whether the teacher has repeated a pupil answer than to decide whether a given question requires the pupil to use higher cognitive processes. Scoring of an earlier set of tapes had indicated that four of the scores were so specific that virtually no differences occurred among raters. Two of these, length of teacher pause and percentage of teacher talk, were obtained by timing teacher pauses and talk with a stop watch. The other two, number of words per pupil response and number of one word responses required a simple word count of each pupil response.

TABLE 2

AVERAGE RELIABILITY FOR THREE GROUPS OF RATERS

Behavior	Group A	Group B	Group C
1. Redirection	.94	.94	.97
2. Number of higher cognitive questions	.62	.57	.59
3. Number of fact questions	.87	.88	.87
4. Prompting	.65	.96	.91
5. Seeking further clarification	.94	.96	.93
6. Repeat questions	.85	.91	.93
7. Answer own questions	.75	.12	.82
8. Repeat pupil answers	.995	.99	.98

Once the initial five tapes had been scored, the remaining tapes were assigned randomly to the raters. When all tapes were scored for the first set of behaviors, raters were trained to score the next set. This cycle was repeated until all tapes had been scored for all behaviors.

RESULTS

One of the twelve behaviors covered in the course, "calling on both volunteers and nonvolunteers" was found to be unscorable. In many classrooms it was not possible to position the camera so all pupils were in camera range, and it was necessary to see the pupil in order to determine whether or not he had volunteered. Two other behaviors, "dealing with incorrect answer in an accepting, non-punitive manner," and "refocusing the pupil's response," were not scored because the inservice field test tapes previously scored (4) showed virtually no variation in these behaviors. Teachers almost never used a punitive manner in dealing with incorrect answers; perhaps because of the presence of the video camera. Refocusing also appeared so rarely in either pre-course or post-course tapes that scoring did not appear worth the equipment and rater time involved.

A total of eleven scores were obtained from each tape. Scores were obtained on nine of the behaviors listed in Table 1, and in the case of one of these, "framing questions that call for longer pupil responses," two scores were obtained. The percentage of teacher talk was obtained by timing pupil and teacher talk with a stop watch. Although not one of the specific behaviors taught, this percentage is an important variable in class discussion and relates to one of the course objectives.

Table 3 gives the pre-course and post-course mean scores for the five groups as well as t ratios between pre and post tape scores for each group. The one-tailed test and the .05 confidence interval were applied in testing the hypotheses.

The reader will recall that Groups A-1 and B-1 were given the entire Minicourse 1; Group A-2 received the entire course except microteaching and videotape feedback, Group B-2 received the entire course except videotape feedback, and Group C-1 received none of the minicourse materials. All subjects were enrolled in student teaching.

The first behavior that Minicourse 1 attempts to change is the length of the student teacher's pause between framing his question and calling upon a pupil to respond. The course attempts to train teachers to pause from 3 to 5 seconds after asking a question and before calling on a pupil. The rationale for this behavior is that such a pause gives pupils more time to frame a thoughtful response to the question. Also, since a pupil's name is not called until after the pause, there is some pressure on all pupils to plan a response in case they are called on. In contrast, calling the pupil's name before or immediately after framing the question immediately notifies other pupils that they are not going to be called upon to respond so they are under no pressure to think through an answer. Of the five groups, only Group B-1 made a significant increase in the average length of this pause.

The course also attempts to increase the number of times student teachers use redirection in the class discussion situation. Redirection is the technique of framing questions in such a way that the question can be directed to several pupils rather than to a single pupil. The teacher asks the question and redirects it to a number of pupils, each of whom contributes to a complete answer. Redirection has the advantage of increasing pupil participation and often leads to direct interaction among pupils in the discussion situation. All groups except C-1 made some improvement in redirection with this improvement reaching statistical significance for groups A-2 and B-1.

Another objective of the course was to train student teachers to ask questions that call for longer pupil responses and to ask fewer questions that can be answered yes or no or with a single word. A word count of pupil responses was made on each tape to determine the average length. It will be noted that the average length of pupil replies on the pre-course tapes ranged from six to seven and a half words. Virtually no change occurred for Group A-2 and Group C between pre-course and post-course tapes. However, significant gains bringing the average length of the pupil reply beyond ten words were made by Groups A-1, B-1 and B-2. These data are supported by previous research on inservice teacher training in which the average length of pupil response was found to increase from 5.63 words to 11.78 words for inservice teachers who were given the complete Minicourse 1 (4). The number of one-word replies was also tallied for 5-minute samples of pre-course and post-course tapes. Groups B-1, B-2, and C all made significant reductions on this variable.

The course also attempted to train student teachers to ask questions that required pupils to use higher cognitive processes in framing their replies. All teacher questions were classified as either fact questions, higher cognitive questions, or procedural questions. The percentages given in Table 3 indicate the proportion of higher cognitive questions to the total of higher cognitive and fact questions combined. It will be noted that there are no significant gains in the percentage of higher cognitive questions and, in fact, a significant loss occurred for Group A-1. In reexamining Group A-1 tapes, it appears that this loss is largely an artifact of the system used for classifying questions. Several student teachers in this group asked questions which called for opinions rather than higher cognitive processes. These were classified as higher cognitive questions, however, on the classification system used. For example, a question such as "How did you like this poem?" is technically an evaluation question although such questions usually elicit replies that do not appear to call for higher cognitive processes. The experienced teachers in the previous inservice sample made less extensive use of opinion questions so that the very simple classification system used seemed adequate. However, in a similar study using teacher interns, Claus (7) found it necessary to set up a separate questioning category for opinion questions which she included along with fact questions and labeled as lower order. Since funds were not available to rescore higher cognitive questions, it will be necessary to await the results of further research before drawing conclusions about the effectiveness of the course in this area. However, it seems doubtful that the course brought about significant increases in the use of higher cognitive questions since such increases would probably have shown up in

TABLE 3

PERFORMANCE OF STUDENT TEACHER GROUPS ON THE BEHAVIORS COVERED IN MINICOURSE 1

	Group 1 - College A						Group 2 - College B						College C		
	Group A-1 (N = 17)			Group A-2 (N = 16)			Group B-1 (N = 15)			Group B-2 (N = 17)			Group C-1 (N = 14)		
Skill or Behavior	Pre Course Mean	Post Course Mean	Pre vs Post t	Pre Course Mean	Post Course Mean	Pre vs Post t	Pre Course Mean	Post Course Mean	Pre vs Post t	Pre Course Mean	Post Course Mean	Pre vs Post t	Initial Mean	Final Mean	Pre vs Post t
1. Length of teacher pause in seconds	1. 28	1. 61	1. 50	1. 37	1. 42	. 35	1. 58	3. 10	2. 17	1. 53	1. 85	1. 45	1. 69	1. 78	. 20
2. Number of teacher redirections	23. 64	26. 52	. 80	23. 12	30. 87	2. 12	19. 93	31. 20	2. 73	25. 05	31. 29	1. 24	20. 78	20. 71	. 02
3. Number of words per pupil response	7. 35	10. 42	1. 94	7. 58	7. 85	. 18	6. 18	10. 13	2. 25	6. 31	10. 73	2. 76	6. 55	7. 43	1. 29
4. Number of one word remarks	4. 11	3. 64	. 32	4. 31	4. 18	. 12	5. 33	2. 40	2. 63	5. 41	3. 47	2. 17	5. 57	3. 64	1. 83
5. Percentage of higher cognitive questions	61. 48	35. 77	2. 82	39. 88	33. 82	1. 14	46. 40	52. 94	. 70	44. 30	44. 06	. 04	27. 08	26. 36	. 11
6. Number of teacher prompts	2. 70	2. 35	. 46	1. 62	2. 31	1. 21	1. 33	1. 86	. 82	1. 05	1. 58	1. 26	1. 28	1. 64	. 54
7. Number of times teacher seeks clarification	3. 47	8. 17	3. 48	4. 68	3. 68	. 89	2. 40	3. 86	1. 65	3. 70	3. 76	. 07	2. 78	2. 57	. 27
8. Number of times teacher repeats question	3. 00	3. 00	00	3. 00	2. 06	1. 08	3. 53	2. 46	1. 28	5. 23	2. 00	3. 67	5. 07	3. 14	1. 91
9. Number of times teacher answers own question	1. 23	. 58	1. 75	2. 50	. 75	2. 43	1. 13	1. 26	. 19	2. 11	. 64	2. 82	3. 01	3. 00	. 00
10. Number of times teacher repeats pupil answer	16. 23	9. 41	2. 18	17. 37	8. 25	2. 60	19. 80	7. 86	4. 89	22. 64	6. 70	4. 66	23. 35	23. 28	. 01
11. Percentage of teacher talk	44. 96	35. 87	2. 48	52. 23	39. 66	2. 36	46. 75	30. 85	4. 00	49. 94	30. 98	4. 52	50. 41	44. 24	1. 91

For groups A-1, A-2, B-2 a t of 1.75 is significant at .05 level. 1.77 is required for Groups B-1 and C-1.

spite of the confounding brought about by inclusion of opinion questions in the higher cognitive category.

Two probing techniques that teachers can use after the pupil's initial response to a question in order to lead the pupil to a more adequate reply are covered in Minicourse 1. These are prompting, in which the teacher gives the pupil clues or follows up his initial response with leading questions and further clarification, in which the teacher attempts to get the pupil to clarify, elaborate or explain his initial response. It may be seen in Table 3 that no significant increases in prompting occurred. Student teachers in Group A-1 made significant gains in their use of further clarification.

Minicourse 1 also attempts to reduce or eliminate the teacher's use of three negative behaviors. These behaviors are repeating the question, answering one's own question and repeating the pupil's answer. Many experienced teachers habitually repeat nearly all of their questions. For example, in the inservice sample previously studied, teachers with an average of 9 years experience repeated their questions nearly fourteen times in a 20-minute discussion lesson. In contrast, it will be noted that the average teacher in preservice training repeated questions five times or less in a 15-minute discussion lesson. Repeating one's questions is considered an undesirable behavior because it takes up valuable discussion time. Also, when used habitually, it conditions pupils not to listen to the first statement of the question. However, there are certainly situations in the typical lesson where it is desirable for the teacher to repeat a question. Since few of the subjects involved in the current research habitually repeated their questions, their behavior for the most part was at an acceptable level prior to the experimental treatment. Two groups, B-2 and C showed significant decreases in this behavior. For Group C this decrease must be attributed to their intervening experience in student teaching. This was probably the main factor operating with Group B-2 as well.

The disadvantages of the teacher answering his or her own questions are obvious. If carried to the extreme, this behavior results in the teacher giving a monologue rather than conducting a discussion lesson. In any case, it deprives pupils of the chance to participate in the discussion and increases the proportion of discussion time taken up with teacher talk. Data from our inservice sample indicated that few teachers answer their own questions with high frequency. This finding was supported in the preservice groups studied where the highest average frequency was three repetitions during a 15-minute lesson. However, in spite of the low initial level of this behavior, significant reductions occurred in Groups A-1, A-2, and B-2.

Many teachers in the inservice study were found to repeat automatically all or most pupil answers. To be classified as a repetition in this study, the teacher had to repeat the pupil's answer nearly verbatim adding no new ideas and making no major changes in the words used. Thus, restatement, elaboration or clarification of the pupil's answer was not classified as a repetition. It may be seen in Table 3 that on the pre-course videotapes, the student teacher groups repeated pupil answers an average of sixteen to twenty-three times in a 15-minute lesson. The frequency of this behavior was reduced significantly in all of the treatment groups. This would suggest that the minicourse, with or without microteaching and videotape feedback, can bring about significant reductions in the frequency of this behavior. It will be noted that virtually no change occurred in this behavior for Group C-1.

One of the objectives of Minicourse 1 is to reduce the proportion of time during the class discussion when the teacher is talking. Previous studies have shown that teachers talk as much as 70 percent of the class time, thereby severely restricting the amount of time available for pupil participation (8, 2). Data from the inservice field test of Minicourse 1 revealed that prior to taking the course, teachers talked an average of 51.64 percent of the time during a class discussion period (4). It may be seen in Table 3 that student teachers also talk a considerable proportion of the time during class discussions. All four of the treatment groups dropped significantly in the percentage of teacher talk. In Group C-1 the percentage of teacher talk was also reduced significantly although a magnitude of this reduction was less than found in the treatment groups. This would suggest that the reduction of teacher talk in the four treatment groups was partially attributable to the student teaching experience and partially to the specific instruction of the minicourse.

SUMMARY AND CONCLUSION

The groups that completed the entire minicourse did not consistently make more or larger changes in behavior than the groups for which some element of the course was omitted. In College A, Group A-1 made significant gains in the desired direction on five of the eleven scores as compared with four significant gains for Group A-2. However, analysis of covariance in which the final performance of student teachers in groups A-1 and A-2 was adjusted for their pre-course performance revealed only a significant difference between the two groups. Group A-2 made significantly greater changes in redirection than Group A-1 (F = 7.66).

In College B, Group B-1, and B-2 both made significant gains in the desired direction in six of the eleven scores. Analysis of covariance revealed no significant differences in the gains made by the two groups.

In College C, three statistically significant gains were made. These gains, however, were generally smaller than those of the treatment groups; none reaching the .01 level of confidence. Thus, Hypotheses 1 and 2 are not supported by the data. The omission of videotape feedback and practice in the microteaching format from the minicourse model appears to have resulted in little or no redirection in the changes in teacher behavior brought about by Minicourse 1.

Hypothesis 3 appears to be supported by the data. The difference in performance between the two groups that completed the entire course, (Groups A-1 and B-1) on one hand and Group C-1 on the other, are fairly large and favor students who completed Minicourse 1. Analysis of covariance revealed that student teachers in groups A-1 and B-1 made significantly greater improvement than Group C in two scores: repeating pupil answers (F=14.85), and percentage of teacher talk (F=5.69).

Interview and questionnaire data obtained from the subjects in the preservice study suggest a number of reasons for the lesser success of the course with these subjects. Probably most significant were the greater demands placed on the preservice teacher as compared with the inservice teacher. Student teachers had a great deal of difficulty carrying out all of the work demanded in the minicourse. Since the course was offered daily, it was necessary for students to do some preparation each evening in order to be ready for the following day's work. This preparation time had to compete with assignments given students in their other classes and with demands made upon them by their supervising teachers and college student teaching supervisors. In some cases, supervising teachers made demands that made it impossible for students in Groups A-1 and B-1 to carry out parts of the minicourse sequence. For example, a number of students were unable to complete all of their practice sessions in the microteaching setting because teachers would not release them at the scheduled time from their duties in the classroom. This experience suggests that in the preservice setting, the course should be offered on a 2 or 3 day per week basis rather than daily. Supervising teachers and other persons having control over the student teacher's time should be very thoroughly briefed on the nature of the minicourse program and the importance of completing all of the required activities.

One surprising outcome of the current study was the somewhat smaller behavioral changes brought about by Minicourse 1 in the presevice setting as compared with those found previously in field testing the course with experienced teachers. For the inservice sample, differences between pre-course and post-course means on all eleven scores obtained in the preservice study were statistically significant. The magnitude of the behavioral changes obtained in the inservice study were also much greater with eight of these differences being significant beyond the .001 level (4).

A final conclusion might be that in spite of a number of problems that arose in conducting Minicourse 1 in the two colleges, several significant changes occurred in the methods of questioning and conducting discussion lessons used by the participating student teachers. Thus, the minicourse model, along with other instructional models that employ microteaching, modeling, and videotape feedback, continues to show promise as a tool for developing specific teacher skills and behavior patterns.

FOOTNOTES

1. The work reported herein was performed pursuant to a contract with the United States Department of Health, Education and Welfare, Office of Education.

REFERENCES

1. Acheson, K. A., "The Effects of Feedback from Television Recording and Three Types of Supervisory Treatment on Selected Teacher Behavior," unpublished doctoral dissertation, Stanford University, 1964.
2. Adams, T. H., "The Development of a Method for Analysis of Questions Asked by Teachers in Classroom Discourse," Doctoral dissertation, Rutgers State University, New Brunswick, New Jersey, 1964.
3. Allen, D. W., and Fortune, J. C., "An Analysis of Microteaching: New Procedure in Teacher Education," in Microteaching: A Description, Stanford University, School of Education, 1966.
4. Borg, W. R., Kallenbach, W. W., Kelley, M. L., and Langer, P., "The Minicourse: Rationale and Uses in the Inservice Education of Teachers," paper presented at the Annual Meeting of the American Educational Research Association, Chicago, February, 1968.
5. Borg, W. R., Kelley, M. L., Langer, P., Kallenbach, W. W., and Gall, M. H., "The Minicourse--Changing Teacher Behavior Through Microteaching," unpublished manuscript, Far West Laboratory for Educational Research and Development, 1968.
6. Bush, R. N. and Allen, D. W., "Microteaching-Controlled Practice in the Training of Teachers," paper presented at the Santa Barbara Conference on Teacher Education of the Ford Foundation, April, 1964.
7. Claus, Karen E., "Effects of Cueing During Modeling and Feedback Sessions on Learning a Teaching Skill," paper presented at the Annual Meeting of the American Psychological Association, San Francisco, August, 1968.
8. Floyd, W. D., "An Analysis of the Oral Questioning Activity in Selected Colorado Primary Classrooms," unpublished doctoral dissertation, Colorado State College, 1960.
9. Hatfield, D. M. and Saunders, B., "Teaching as Others See Us," Educational Instructional Broadcasting, pp. 27-29, April, 1968.
10. Kallenbach, W. W., and Gall, M. D., "The Comparative Effects of Microteaching and a Conventional Training Program on the Classroom Performance of Elementary Intern Teachers," unpublished manuscript, Far West Laboratory for Educational Research and Development, 1968.
11. Orme, M. E., "The Effects of Modeling and Feedback Variables on the Acquisition of a Complex Teaching Strategy," unpublished doctoral dissertation, Stanford University, 1966.
12. Tuckman, B. W., and Oliver, W. F., "Effectiveness of Feedback to Teachers as a Function of Source," Journal of Educational Psychology, pp. 297-301, August, 1968.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

LEVELS OF COGNITIVE FUNCTIONING AS RELATED TO ANXIETY[1]

JERRY S. CARLSON
University of California, Riverside

FRANK L. RYAN
University of Minnesota

ABSTRACT

The relationships between levels of cognitive functioning as suggested by the *Taxonomy of Educational Objectives*, *Handbook I* and states of affective arousal as measured by Text Anxiety (the main variable of interest), General Anxiety, and Need for Achievement were investigated. Two separate samples were used: (1) 234 university students, and (2) 319 junior and senior class high school students.

For the university sample significant negative correlations between Text Anxiety and Knowledge (recall of specific facts) and Test Anxiety and Comprehension were found. For the high school sample the negative correlation between Test Anxiety and Comprehension was significant. No significant relations between Test Anxiety and higher levels of cognitive performance were found. The results are discussed in terms of interfering effects to task performance that Test Anxiety may produce when the examinee is given no information to "workwith" and must rely essentially on memory.

THE FACT that a relationship exists between affective states and cognitive functioning has been well established (6, 2, 10). Evidence indicates that anxiety can act as a motivational force on test performance; but if the difficulty of the task is increased, those individuals with high test anxiety do not do as well as those with low test anxiety (5). Sarason and others (9) have suggested that test anxiety seems to have an "interfering effect" and can act as an intervening variable on problem-solving performance.

Sarason (8) demonstrated that highly test anxious individuals differ from those with low anxiety in that the former emit responses that are detrimental to the successful completion of the task. According to Sarason, high anxiety states manifest themselves in two ways: (a) self-verbalization ("I can't pass this test"); and (b) autonomic responses which are picked up by higher GSR ratings. Accordingly, highly anxious persons emit more interfering responses than do those individuals who exhibit less anxiety (4, 9). Certain levels of anxiety, however, can be optimal to test performance and levels either above or below this result in impaired test performance. Experimental support that the relationship between anxiety level and test performance can be described by an inverted U has been offered by Stennet (11) and Korchin (3). The variable of performance generally considered in anxiety-test performance relationships is level of task difficulty.

Moving to a consideration of the interaction of level of anxiety and performance with different problem types, Speilberger (10) reported that highly anxious subjects did as well as subjects who exhibited lower anxiety on clearly worded problems that suggested an answer rather quickly. Performance decrements, however, for the highly anxious group were noted when the protocols were ambiguous or unclear. It was concluded that "task-produced interfering response tendencies interacted with high drive to bring about the observed effects" (10:372).

In summary, there is evidence that levels of affective arousal are related to the individual's ability to solve problems of various types and degrees of difficulty. The present investigation is concerned with further examination of such relationships, although from a perspective that delineates test taking behavior into specific levels of cognitive functioning. The *Taxonomy of Educational Objectives* (1) provides rubrics for the classification of cognitive functioning in a test situation that are (1) independent from one another, and (2) represent a hierarchy of levels.

PURPOSE

The purpose of the present investigation was to ascertain the relationships, if any, between anxiety and performance at different levels of cognitive functioning.

Experiment I

METHOD

Evaluation Instruments

Four evaluative instruments were constructed to reflect four of Bloom's (1) classifications for cognitive behavior: Knowledge, Comprehension, Application, and Analysis. The content of the four multiple-choice instruments developed was taken from the physical and biological sciences. To ascertain the appropriateness of the test items (e.g. level of difficulty, clarity of wording, and ability to discriminate) a pilot try-out of seventy-eight of the "initially generated" items was conducted. From this it was possible to create helpful guidelines for further item writing and selecting.

Four individuals with familiarity and experience in using Bloom's taxonomy, independently assigned each of the resultant test items to one of the four cognitive classifications being considered in this study. Item classifications not agreed upon by at least three of the four persons were either rewritten to become "acceptable" or were omitted. The final group of test items employed was selected through item analysis procedures. Split-half reliabilitites of the four evaluation instruments, corrected with the Spearman-Brown Prophecy Formula (12) were: Knowledge, .81; Comprehension, .76; Application, .85; and Analysis, .61.

SAMPLE

The sample consisted of 234 students (142 women, 92 men) enrolled in Education or Western Civilization courses at the University of California, Riverside. Junior, senior, or fifth-year standing is required for admission to Education courses. Western Civilization is a lower division course required of all students, with the large majority of students completing the requirement in either their freshman or sophomore year. Thus, the sample was judged as generally representative of the population of students at the University of California, Riverside.

Procedures

All Ss were randomly assigned to one of four groups: Knowledge (K); Comprehension (C); Application (Ap): and Analysis (An). All Ss were administered the Test Anxiety (TA), General Anxiety (GA), and Need for Achievement (NA) sections of the Autobiographical Survey developed by Sarason (7). Test-retest reliabilities above .78 for TA and GA, and above .70 for NA have been reported (Sarason, 1958). Evidence of the equivalence of the groups was offered through analysis of variance for each of three independent variables: TA ($F=0.33$; d.f. $=3/233$; N.S., $p < .01$), GA ($F=0.51$; d.f. $=3/233$; N.S., $p < .01$), and NA ($F=1.38$; d.f. $=3/233$; N.S., $p < .01$).

One week after taking the tests described above, all Ss were given the science test appropriate to their group (i.e. K, C, Ap, or An). No time limits were imposed for any of the tests.

RESULTS

Correlational analysis was performed on the total test scores for each group. The assumption of linearity was confirmed by visual inspection of the graphed data. The correlation matrix for the group receiving the Knowledge science test (N=59) is reported in Table 1. Positive correlation coefficients departing significantly from zero were between General Anxiety and Test Anxiety ($p < .01$); and Test Anxiety and Need for Achievement ($p < .05$). The negative correlation between Test Anxiety and Knowledge was significant ($p < .05$).

TABLE 1

MATRIX OF INTERCORRELATIONS FOR KNOWLEDGE GROUP

Variable	1	2	3	4
1. Test Anxiety	-			
2. General Anxiety	.63**	-		
3. Need for Achievement	.28*	.11	-	
4. Science Knowledge	-.27*	-.21*	-.13	-

* $p < .05$
** $p < .01$

Results of analysis of the data from those subjects who took the Comprehension science test (N=58) are shown in Table 2. Significant positive correlations were found between TA and GA ($p < .01$); and GA and NA ($p < .05$). Significant negative correlations were between TA and Comprehension ($p < .01$); and GA and Comprehension ($p < .05$).

TABLE 2

MATRIX OF INTERCORRELATIONS FOR COMPREHENSION GROUP

Variable	1	2	3	4
1. Test Anxiety	-			
2. General Anxiety	.63**	-		
3. Need for Achievement	.19	.32*	-	
4. Science Comprehension	-.43**	-.29*	-.07	-

* $p < .05$
** $p < .01$

Table 3 is the correlation matrix of the scores obtained from the group which received the Application science test (N=62). Significant positive correlation coefficients were found between GA and TA ($p < .01$); NA and GA ($p < .01$); and NA and GA ($p < .01$).

TABLE 3

MATRIX OF INTERCORRELATIONS FOR APPLICATION GROUP

Variable	1	2	3	4
1. Test Anxiety	-			
2. General Anxiety	.64**	-		
3. Need for Achievement	.60**	.51**	-	
4. Science Application	-.17	-.03	-.17	-

** $p < .01$

The only significant correlation for the group which received the Analysis science test (N=55) was positive and between TA and GA ($p < .01$). Table 4 reports the correlation matrix.

TABLE 4

MATRIX OF INTERCORRELATIONS FOR ANALYSIS

Variable	1	2	3	4
1. Test Anxiety	-			
2. General Anxiety	.67**	-		
3. Need for Achievement	.21	.14	-	
4. Science Analysis	-.02	.09	-.11	-

** $p < .01$

DISCUSSION

The purpose of this investigation was to examine the interaction of various states of affective arousal, as measured by GA, TA and NA, and level of cognitive functioning (Knowledge, Comprehension, Application, and Analysis). Hence, the intercorrelations found between GA, TA and NA for all groups tested is of no particular interest; although one obvious conclusion is they do not seem to be measures of independent affective states. Although there appears to be a lack of consistency of correlations between TA and NA (significantly different from zero for the Knowledge and Application groups); GA and NA (significantly different from zero for the Comprehension and Application groups), such is not the case. Intercorrelations between all groups were not significantly different from one another: TA and GA ($X^2=.163$; d.f.=3; N.S., $p < .01$); GA and NA ($X^2=6.85$; d.f.=3; N.S., $p < .01$); TA and NA ($X^2=8.83$; d.f.=3; N.S., $p < .01$).

The significant negative intercorrelations found between TA and Knowledge, TA and Comprehension, and GA and Comprehension indicate that states of anxiety are differentially related to levels of cognitive functioning. That is, higher levels of TA have a debilitating effect on the "lower" but not the "higher" levels of cognitive functioning. The significant negative correlation between GA and Comprehension probably appeared because of the lack of independence between GA and TA.

The findings of Speilberger (10) which indicated that Ss with high and low anxiety function at the same level when unambiguous and clearly worded problems are presented provide a possible analogy to the results of this study. Although the protocols used were, hopefully, clear and unambiguous, the types of questions asked and the information given for each of the four levels of Bloom's (1) taxonomy used provides us with the basis for our analysis.

The items for the Knowledge test simply asked the S to remember some scientific facts. No communication beyond the question was given. The same is true, though to a lesser degree, for the Comprehension test. The Application and Analysis tests provided a communication with information that allowed the S to "work with" the item and "think it out." For intelligent individuals (UC draws its student body from the upper 12 percent of high school graduates) this may have been a much easier and less frustrating task than having to simply recall factual information. The Knowledge and Comprehension questions may have provided stimuli for those individuals with high TA and GA, which resulted in their making task irrelevant responses which led to lower performance than that shown by the low anxiety Ss. At any rate, it appears that the interfering responses generated were differentially developed by those of high and low anxiety and the type of question, which implies levels of cognitive functioning required to give an answer, played a crucial role in test performance.

Experiment II

Whether or not the relationships between levels of cognitive functioning and test anxiety found for the university student sample used in Experiment I exists for individuals of less academic ability and less sophistication and experience with test taking was the research question of the second experiment.

METHOD

Evaluation Instruments

The four evaluative instruments utilized in Experiment I were renormed for the Ss of Experiment II.

The following reliabilities (Kuder-Richardson, Formula 20) were calculated: Knowledge, .73; Comprehension, .73; Application, .81; Analysis, .72.

Sample

The sample consisted of 319 junior and senior class students from the high school of a unified school district. STEP-SCAT data indicated a sample which was academically average. The school serves a middle to lower middle SES community.

Procedures

As in Experiment I, Ss were randomly assigned to one of four groups: Knowledge, Comprehension, Application, or Analysis. One week after taking the Test Anxiety part of Sarason's Autobiographical Survey (7) the Ss were administered the evaluation instrument appropriate to their group. Analysis of Variance showed the four groups did not differ significantly on the independent variable, TA (F=0.15; d.f.=3/315; N.S., $p < .01$).

RESULTS

Correlational analysis showed a significant negative correlation between TA and level of cognitive functioning for the group which received the Comprehension science test ($r = -.32$, $N = 83$, $p < .01$). No other correlations were significant. The results are reported in Table 5.

TABLE 5

CORRELATIONS BETWEEN TEST ANXIETY AND LEVEL OF COGNITIVE FUNCTIONING

Level of Cognitive Functioning	N	Correlation with TA
Knowledge	83	.01
Comprehension	83	-.32**
Application	70	.12
Analysis	70	.13

** $p < .01$

SUMMARY

This study was designed to explore the relationship between levels of cognitive functioning and states of anxiety. Levels of cognitive functioning were designed to reflect the Knowledge, Comprehension, Application, and Analysis parts of Bloom's Taxonomy of Educational Objectives, Handbook I (1). States of anxiety were measured by the Autobiographical Survey (7). The variable of interest was Test Anxiety. Experiment I, using a sample of "bright" and test sophisticated university students, revealed significant negative correlations between Test Anxiety and Knowledge ($p < .05$), and Test Anxiety and Comprehension ($p < .01$).

The study was extended to a second part in order to ascertain if the relationships established above held for a younger, generally less intelligent, and less test sophisticated population. Hence, in Experiment II, a sample of junior and senior class high school students of average ability from a middle to lower middle SES area was chosen. Procedures generally replicated those employed in Experiment I. The results indicated a significant negative correlation ($p < .01$) between Test Anxiety and Comprehension. Thus the findings of Experiment I were partially replicated.

FOOTNOTE

1. The research reported here was partially supported by an intramural research grant, University of California, Riverside.

REFERENCES

1. Bloom B. S. and others, Taxonomy of Educational Objectives, Handbook I: Cognitive Domain, David McKay, New York, 1956.
2. Harleston, B. W., "Test Anxiety and Performance in Problem-Solving Situations," Journal of Personality, 30:557-573, 1962.
3. Korchin, S. J., "Anxiety and Cognition," in C. Scheerer, ed., Cognition Theory, Research, Promise, Harper and Row, New York, 1964.
4. Mandler, G., and Sarason, S. B., "A Study of Anxiety and Learning," Journal of Abnormal and Social Psychology, 47:166-173, 1952.
5. Pickerel, E. W., "The Differential Effect of Manifest Anxiety on Test Performance," Journal of Educational Psychology, 49:43-46, 1958.
6. Ruebush, B. K., "Interfering and Facilitating Effects of Test Anxiety," Journal of Abnormal and Social Psychology, 60:205-212, 1960.
7. Sarason, I. G., "Effects on Verbal Learning of Anxiety, Reassurance, and Meaningfulness of Material," Journal of Experimental Psychology, 56:472-477, 1958.
8. Sarason, I. G., "Intellectual and Personality Correlates of Test Anxiety," Journal of Abnormal and Social Psychology, 59: 272-275, 1959.
9. Sarason, S. B. and others, "The Effect of Differential Instructions on Anxiety and Learning," Journal of Abnormal and Social Psychology, 47: 561-565, 1952.
10. Spielberger, C. D., "The Effects of Anxiety on Complex Learning and Academic Achievement," in C. D. Spielberger, ed., Anxiety and Behavior, Academic Press, New York, 1966.
11. Stennet, R. G., "The Relationship of Performance Level to Level of Arousal," Journal of Experimental Psychology, 54: 54-61, 1957.
12. Thorndike, R. L., and Hagan, E., Measurement and Evaluation in Psychology and Education, John Wiley and Sons, Inc., New York, 1961.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

CONCEPT-SWITCHING TASK PERFORMANCE AS A FUNCTION OF REWARDS

PHILIP C. CHINN,
The University of Utah

OLIVER P. KOLSTOE
Colorado State College

ABSTRACT

The Kounin (4) concept-switching task was administered to groups of institutionalized mentally retarded, non-institutionalized mentally retarded, lower-class normal, and middle-class normal children matched on MA. The concept switching task was administered under three different reward conditions: promise of reward, low tangible reward, and high tangible reward. Differences in performance between groups was tested by means of a 3 x 4 analysis of variance. F tests yielded no significant differences at the .05 level. The study failed to confirm either Kounin's concept of rigidity or Zigler's deprivation motivational hypothesis. Possible reasons for lack of confirmation are discussed.

EVER SINCE Lewin (5) postulated that the difference between a feebleminded and normal child of the same degree of psychical differentiation (mental age) is in the smaller capacity for dynamic rearrangement in the psychical systems of the feebleminded, the notion of rigidity has intrigued investigators. Kounin (4) offered the results of five experiments to demonstrate the tendency of the mentally retarded person to cling to a fixed goal or habit thus making it difficult for him to evaluate new situations.

Studies by Zigler and Unell (11), Zigler and deLabry (9), and Zigler, Hodgen, and Stevenson (10), however, suggest that incentive variables significantly affect experimental performance. Additionally, Heber (3) reported that an incentive which is highly preferred by the individual being tested seems to produce superior responses regardless of intelligence level.

Investigations of the effects of institutionalization and social class on various incentive responses have been limited, but provocative. Terrell, Durkin, and Wiesley (7) found that middle-class normal children seem to desire to perform correctly for the sake of being correct regardless of the incentive offered. Lower-class individuals have tended to function at their best levels when a tangible or material incentive is offered as compared to an intangible incentive. According to Zigler and deLabry (9), institutionalized mentally retarted individuals and lower-class individuals matched on mental age have performed comparably under the same incentive conditions.

It is Zigler's contention that institutionalized subjects tend to prolong adult interaction by continuing their tasks as long as possible. This is interpreted to reflect motivation rather than rigidity. In addition, Zigler (8), Backer (1), and Zigler, Hodgen, and Stevenson (10) found evidence that the institutionalized subjects tend to make few errors on their tasks. They consider this also due to a high motivation to secure adult approval.

A study by Terrell (6), while it did not deal with retarded subjects, included a promise of reward. The promise of a reward, when compared with more tangible rewards, was less motivating for the subjects.

Zigler (12) has proposed that the lack of psychic flexibility of the mentally retarded may not be due to rigidity as suggested by Lewin, but may rather be due to social deprivation (lack of adequate affectional adult contact). If this is correct, one would expect mental retardation, institutionalization, and social class to effect differential performance under different reward conditions when the performance task requires cognitive flexibility.

This notion has been investigated by three studies. In addition to the original study by Kounin (4) on concept-switching, two other studies have used his same materials and tasks. Zigler and Unell (11) compared the performance of a group of institutionalized retardates with a group of normals using a nonreinforcement condition and a reinforced condition.

They found that both groups performed better under tangible reinforcement, but the normal subjects did no better than the retarded regardless of the reinforcement. This they interpret as supporting the motivational hypothesis.

Zigler and deLabry used the concept-switching task of Kounin with an institutionalized mentally retarded group, a group of lower class normals, and a group of middle class normals. They used two reward conditions: intangible (being correct) and tangible (choice of a toy). Their principle finding was that each group performed best under optional reward conditions; intangible for the middle-class normals and tangible for the mentally retarded and the lower-class normals. This evidence tends to support the motivational hypothesis of Zigler.

STATEMENT OF THE PROBLEM

The purpose of the present study was to compare the performance of four groups of subjects of comparable mental ages on the concept-switching task used by Kounin (4) under three different reward conditions. The groups were:

1) Institutionalized educable mentally retarded.
2) Non-institutionalized educable mentally retarded.
3) Middle-class normals.
4) Lower-class normals.

The reward conditions were:

1) Promise of a reward.
2) Low tangible reward.
3) High tangible reward.

METHOD

The subjects for the present study included thirty non-institutionalized mentally retarded children selected from special education classes at the Colorado State College Laboratory School and the Waco Independent School District, Waco, Texas. Thirty institutionalized mentally retarded subjects were selected from the Mexia State School in Mexia, Texas. A total of sixty normal subjects were selected from elementary schools in the Waco Independent School District. Thirty normal subjects were chosen from middle-class socio-economic backgrounds as determinded by a modified version of the Warner's Index of Social Characteristics (2), and thirty normal subjects were from lower-class socio-economic backgrounds as determined by the same Index.

TABLE 1

MEAN CHRONOLOGICAL AGES (IN MONTHS), AND IQ'S OF ALL EXPERIMENTAL GROUPS

Groups		Inst. EMR	Non-Inst. EMR	Lower-Class Normal	Middle-Class Normal
Promise of Reward	CA	162.6	135.2	95.2	88.4
	MA	90.9	84.2	98.7	93.5
	IQ	65.2	68.5	96.1	102.9
Low Tangible Reward	CA	155.7	148.1	103.2	91.9
	MA	85.9	85.9	95.3	91.1
	IQ	63.6	66.4	95.8	99.3
High Tangible Reward	CA	167.6	145.0	91.2	93.9
	MA	88.7	87.9	97.6	98.6
	IQ	63.4	67.0	105.2	102.6

NOTE: CA--Chronological Age, MA--Mental Age, IQ --Intelligence Quotient

All subjects were administered the Peabody Picture Vocabulary Test to determine mental age and IQ scores. The characteristics of the groups are given in Table 1.

Kounin's concept-switching task (4) was administered to determine the subjects' abilities to switch concepts under three reward conditions. The reward conditions were: 1) a high tangible reward of candy, 2) a low tangible reward of a slip of paper with which to redeem an unknown reward later, 3) and a promise of a reward. The description of the concept-switching tasks and the procedures following the administration are as follows: Task I, a color grouping, and Task II, a form grouping, were pre-experimental. These tasks gave the subjects experience with two principles of classification and indicated whether or not they could make the appropriate form and color discriminations. Task III was the experimental task involving either a color or a form grouping.

Task I.--Twenty-five colored rectangular cards, measuring 3.5 x 2.5 inches, were presented to each subject who was asked to classify them into groupings. There were five each of the following colors: orange, yellow, green, blue, and red. The instructions for this and subsequent tasks were similar to those used by Kounin:

> Here are some things all mixed up. Some of these are the same and belong together. Here is a box with five sections. Suppose you find some that are the same and belong together and you put them in one of the sections. Then find some others that are the same and put them in another section, and keep going until you have all those that are the same together.

Task II.--Twenty-five white cards were used. There were five cards of each of the following shapes: triangles, squares, circles, crosses, and 5-pointed stars. These were presented to each subject who was again required to classify them into five groupings.

Task III.--Twenty-five cards were used whose colors were the same as those used in Task I and whose forms were the same as those used in Task II. Again, there were five of each type of form, with each of the same form painted a different color, i.e., an orange, yellow, green, blue, and red triangle, etc. Each color was represented in each shape once. This task had two parts:

Part 1.--The twenty-five cards were presented to each subject, and he was asked to classify them into five groupings.

Part 2. --After the subject had classified the cards correctly, on the basis of either color or form, the experimenter reshuffled the cards, returned them to the subject, and asked him to put them together some other way.

PROCEDURE

Each subject was administered the experimental task individually. At no time was any form of verbal praise offered except in relation to the reward the subject was attempting to earn. The subject was told:

> We're going to play some games today and give you a chance to win some prizes if you want to. Do you want to? You don't have to if you don't want to. Are you sure you want to?

The subject was initially given one of the pre-experimental tasks, either Task I or Task II. After completing this task, he was given the other pre-experimental task. One-half of the subjects in each group was given Task I followed by Task II, with the other half being given Task II followed by Task I. The subjects were given Part 1 of Task III, followed by Part 2 of Task III.

In the high tangible condition (choice of several 5 to 10 cent items such as gum, candy, and Cracker Jacks), when the subject was seated, he was told:

> We're going to play some games in which you can win a prize. Every time you get a game right, you can choose any one of these prizes. (The subject was shown the items mentioned above which were placed on the table.)

Upon successfully completing Task I, Task II, and Part 1 of Task III, the subject was allowed to take his choice of the items on the table. Prior to Part 2 of Task III, which required the subject to switch concepts, the subject was told:

> See if you can win a prize by putting them together some other way.

If he did not switch on the first trial, he was then told:

> You didn't win a prize that time. See if you can put them together some other way.

If the subject still did not change his method of grouping, he was instructed as follows:

> No, you can win a prize by finding some other way that they may belong together.

These latter instructions were continued until the subject had a maximum of seven total trials or a correct grouping, either of which terminated the experiment for the subject. If the subject hesitated in continuing during any of the seven trials, he was told:

> Sure you can do it. Remember, you can win a prize if you can find some other way to do it.

If the subject did not switch concepts by the seventh trial, he was terminated, and was automatically given a score of eight.

In the low tangible condition, the subjects were given the same instructions, but instead of showing them the items on the table, they were shown a stack of slips of paper, marked "One prize (1)" lying on the table. If a subject was unable to read the slip of paper, it was read to him. As he successfully completed each part, he was given a slip of paper. The words "slip of paper" were substituted for "prize" throughout the experiment using the low tangible condition.

In the promise of reward condition, the subject was not shown anything. The initial instructions were:

> We're going to play some games. For every game you get right, you will receive a prize after everyone else has had a chance to play the games. We will play four games, so if you get all four right, you will receive four prizes.

The remainder of the instructions were the same as for the high tangible condition, with the comment after each successful part:

> That was right. After everyone else has a chance to play the game, you will get a prize for that game.

Since the promise of reward condition groups were tested first, these subjects had no knowledge of the nature of the prize for which they were actually working. The low tangible group was tested after the promise condition was completed, so that this group had only the slip of paper which they were told they would be able to redeem for a prize, the nature of which was unknown to them.

Because of the need for sequencing the reward conditions from promise to low tangible to high tangible, subjects were assigned to conditions on the basis of the order in which they were identified: the first ten S's of group 1 to condition 1, and the next 10 S's of group 1 to condition 2, etc.

FINDINGS AND DISCUSSION

The groups were tested for differences by means of a 3 x 4 factorial analysis. In testing for an interaction component, an F test was utilized. The F test resulted in a value of 1.146 which was not significant at the . 05 level of confidence.

An F test was utilized to determine significance of main effects involving the four groups. This resulted in a value of .045 which was not significant at the . 05 level of confidence.

An F test was utilized to determine significances of main effects involving the three reward conditions. This resulted in a value of 2. 122 which was not significant at the . 05 level of confidence.

Of the 120 subjects in the experiment, eight failed to complete the task of concept switch. Multiple regression analysis was therefore used to examine for possible differences between subjects in the four groups and under the three reward conditions using only those 112 subjects who were successful in accomplishing the switch. Using number of trials to completion as the criterion, cell membership in the 3 x 4 design and IQ as predictor variables yielded a multiple R^2 of .075 and a non-significant F of .74.

Table 2 is a summary table presenting the statistical treatment of the number of trials to completion on the concept switching task for all subjects in the experiment.

TABLE 2

ANALYSIS OF VARIANCE FOR TRIAL TO COMPLETION

Source	Sum of Squares	DF	M-Square	F
Columns (Rewards)	0.3166	2	0.1583	0.045
Rows (Groups)	22.2916	3	7.4305	2.122
Interaction	24.0833	6	4.0138	1.146
Within Groups	378.1000	108	3.5009	
Total	424.7916	119		

DISCUSSION

The data from the present study failed to support either Kounin's theory of psychic rigidity of Zigler's differential motivational hypothesis concerning institutional and social deprivation. This was true whether all the subjects were included in the analysis or whether only those who successfully performed the concept switching task were compared. Just why this empirical test failed to substantiate either Kounin's theory or Zigler's hypothesis is not clear. Several possibilities can be considered.

First, the subjects employed in the present study differed considerably from those employed by other investigators. The subjects used in this study had mental ages about a year and a half older than those subjects studied in previous research. Likewise, the chronological ages represented in this study were from 1 to 2 years superior for the retarded subjects and from 2 to 4 years for the normal. Thus, either mental maturity and/or chronological maturity might account for the findings.

Second, it is possible that since the testing occurred in school settings where the students had been for some time, that a promise of reward, a slip of paper believed to be redeemable for a reward and an actual immediate prize were perceived by the subjects as comparable motivating factors. That is, the school experiences of the subjects may have been such that they learned to believe that school associated adults keep their promises.

The findings of the study and the discussion suggest:

1. Kounin's theory of rigidity and Zigler's differential motivational hypothesis are not tenable.
2. Both the theory and the hypothesis are tenable, but are modified as a function of age and/or mental age.
3. Both the theory and the hypothesis are tenable but are modified as a function of experience and/or training.

The need for further research is self-evident. Although studies utilizing designs similar to the present study could use populations similar to those of the earlier studies, it would seem somewhat more efficient to employ a complex design which might allow for the identification of possible interaction effects.

In any case, the implications for classroom precedures are much too important to be ignored. If indeed tangible, low tangible and promises of reward are equally motivating, classroom reinforcement can be greatly simplified. Similarly, if response to various types of reinforcement can be altered by experience or is altered by maturation, this may have an effect on classroom procedure.

Conversely, should future research sustain the Kounin theory or the Zigler hypothesis, this may generate a reexamination of this important issue.

REFERENCES

1. Backer, M. H., "An Experimental Investigation of the Motivational Hypothesis of Rigidity in Retardates: A Comparison of Retardate and Normal Performance of a Series of Card Sorting Tasks," Dissertation Abstracts, 26:4068-4069, 1966.
2. Bransford, L. A., "A Comparative Investigation of Verbal and Performance Intelligence Measures at Different Age Levels with Bilingual Spanish Speaking Children in Special Classes for the Mentally Retarded," unpublished doctoral dissertation, Colorado State College, 1966.
3. Heber, R., "Motor Task Performance of High Grade Mentally Retarded Males as a Function of the Magnitude of the Incentive," American Journal of Mental Deficiency, 63:667-71, 1959.
4. Kounin, J. S., "Experimental Studies of Rigidity: I. The Measurement of Rigidity in Normal and Feebleminded Persons," Character and Personality, 9:251-273, 1941.
5. Lewin, K., A Dynamic Theory of Personality, McGraw-Hill, New York, 1935.

6. Terrell, G. Jr., "The Role of Incentives in Discrimination Learning in Children," Child Development, 29: 231-236, 1958.

7. Terrell, G. Jr., Durkin, K., and Wiesley, M., "Social Class and the Nature of the Incentive in Discrimination Learning," Journal of Abnormal and Social Psychology, 59: 270-272, 1959.

8. Zigler, E. F., "Social Deprivation and Rigidity in the Performance of Feebleminded Children," Journal of Abnormal and Social Psychology, 57: 413, 1961.

9. Zigler, E. F. and deLabry, J., "Concept Switching in Middle Class, Lower Class, and Retarded Children," Journal of Abnormal and Social Psychology, 65: 267-268, 1962.

10. Zigler, E. F., Hodgen L., and Stevenson, H., "The Effect of Support and Non-support on the Performance of Normal and Feebleminded Children," Journal of Personality, 26:106-122, 1958.

11. Zigler, E. F. and Unell, E., "Concept Switching in Normal and Feebleminded Children as a Function of Reinforcement," American Journal of Mental Deficiency, 66:651-657, 1962.

12. Zigler, Edward, "Research on Personality Structure in the Retardate," in Ellis, Norman R. (Ed.), International Review of Research in Mental Retardation, Academic Press, New York, 1966.

PERTINENT RECENT TITLES FROM DERS

THE MEASUREMENT AND PREDICTION OF TEACHER EFFECTIVENESS
By Prof. A. S. Barr et al.
144 pages. $5.

Can good teachers be distinguished from poor teachers? This milestone monograph presents an overview of 30 years of important research and offers some general observations and new hypotheses.

ABILITY GROUPING IN THE PUBLIC SCHOOLS
By Prof. Walter R. Borg
104 pages. $3.

This significant field study analyzes the generally superior learning achievements of elementary, junior high, and high school pupils in an ability-grouping system, and discusses accompanying personality problems.

YEAR-ROUND EDUCATION
By Prof. Clarence A. Schoenfeld and Neil Schmitz
144 pages. $3.

The problems and possibilities in the all-year calendar, from kindergarten to college. A unique assessment of a national issue that focuses on facts rather than on fancy.

SCHOLARS GUIDE TO JOURNALS OF EDUCATION
By Profs. L. Joseph Lins and Robert A. Rees
160 pages. $3.95.

An annotated index of 135 leading journals of education and educational psychology so authors and teachers can find out what's printed and how to communicate their findings.

ders

DEMBAR
EDUCATIONAL
RESEARCH
SERVICES, INC.

POST OFFICE BOX 1605 . MADISON, WISCONSIN 53701

PLEASE SEND ME THE INDICATED NUMBER OF COPIES OF:

☐ The Measurement and Prediction of Teacher Effectiveness at $5.
☐ Ability Grouping in the Public Schools at $3.
☐ Year-Round Education at $3.
☐ Scholars Guide to Journals of Education at $3.95.

Name ______________________________
Address ______________________________
City ______________________________
State ______________ Zip ___________
☐ I enclose a check for postpaid books.
☐ Bill me and I'll pay the postage.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

TREATMENT COMPARISONS: ITEM RESPONSES IN MULTI-FACTOR REPEATED MEASUREMENTS DESIGNS[1]

RAM K. GUPTA

University of Alberta,
Edmonton Canada

ABSTRACT

Scores on an objective test representing the number of correctly answered items should not be used as a criterion for judging the relative efficacy of 'treatments' whenever item-treatment interaction is present. For valid comparisons of treatments, one needs items which are affected by but do not interact with the treatments. Since this condition cannot be satisfied through *a priori* means, the researcher has to approach it *a posteriori*, analyzing his data to see whether interaction is present. The analyses involve the application of repeated measurements designs in which each item functions as a criterion and constitutes a level of the factor containing repeated measures. The design would contain the 'treatments' as another factor—it may also contain additional factors. The above premise is substantiated by the results from a four factor design, containing repetitions on one factor.

PUPIL SCORES on objective tests are often used as a basis for judging the relative efficacy of educational innovations or 'treatments,' such as different curricula, instructional devices, or teaching methods. The scores are usually analyzed in factorial designs at various levels of complexity. However, the scores representing the number of correctly answered items could be too gross a criterion for drawing valid conclusions regarding the efficacy of 'treatments.' This was demonstrated by Moonan (9:253) in this Journal through the following example:

> Suppose a fifty item test is given to two classes... of arithmetic students who were taught by two different methods. If the students of the first class answered the first 25 items correctly and missed the last 25 items, while the students of the second class reversed this type of response, then the two sample means which are calculated from the total test scores of the two groups would be equal and the hypothesis of equality of the two methods would be sustained.

In the above example, the methods under comparison would prove to be equally effective when judged on the basis of total scores. On the basis of the items in the first half of the test, however, the first method would prove to be superior, the reverse being the inference from the items in the second half of the test. The contradiction arises from the presence of interaction between the 'treatments' and the (items belonging to the) two halves of the test. Moonan recommended, therefore, that "we need to use an examination with items which are affected by, but do not interact with, the factors of experiment."

Since a researcher cannot ensure *a priori* that the test items he plans to use will not interact with the "factors of experiment," he has to approach the problem *a posteriori*. After having selected the test on the basis of other relevant considerations and using it, he can analyze the data to find out whether the items of the test interacted with the 'treatments' and with other "factors of experiment," if present. Such analyses involve the application of repeated measurements designs, such that every item works as a criterion and constitutes a level of the factor containing repeated measures. Such a design would contain at least one more factor—that pertaining to the 'treatments' being compared. It may also include additional factors or variables which the researcher may like to examine simultaneously.

The various factors of a repeated measurements design can be visualized through a two-way layout in which the column headings represent test items furnishing repeated measures and the row headings stand for the various groups of individuals answering the items.

The simplest design would contain only two factors, one having repeated measures and the other 'treatments.' In such a design, the subjects are classified only in terms of the 'treatments' administered to them. In the three factor design with repetitions on one factor, the subjects are also classified on the basis of an additional variable, for example, sex. In the four factor design of this type, classification is based upon yet another variable, such as the socio-economic status or ability levels of the subjects, and so on. Similarly, instead of a single factor containing repetitions, one can use two or more such factors.

Irrespective of the number of factors included in a repeated measurements design, the analysis of data contains two parts:

(a) a 'between subjects' or 'between subjects within groups' part in which the total scores constitute the criterion. As such, the hypotheses examinable in this part of the analysis are called here 'score' hypotheses. Most researchers limit their analysis to testing such hypotheses only.

(b) a 'within subjects' part in which each test item furnishes a criterion. Hence the hypotheses testable under this part are termed here 'item' hypotheses.

The present paper aims at showing the limitations inherent in confining attention to 'score' hypotheses and at exemplifying the contribution which an examination of 'item' hypotheses can make to the evaluation of 'treatments.'

METHOD

The design of the experiment used in this research had four factors, of which three were classificatory in nature. Factor A consisted of fifteen mathematics teachers in the U. S. Midwest who voluntarily participated in the study. Factor B represented the students' sex. Factor C denoted two types of textbooks--'SMSG' and 'conventional.' Each teacher had taught concurrently two sections of a ninth grade, one using an 'SMSG' text developed by the School Mathematics Study Group (1), and the other section using any of the various 'conventional' texts.

There were 513 subjects—232 girls and 281 boys. The SMSG text was used by 255 of them and the 'conventional' texts by 258. Two-thirds of the teachers had randomly assigned the students to the two types of texts.

The data consisted of the subjects' responses to the Sequential Tests of Educational Progress, Mathematics, Form 2A (STEP-Math 2A) published by the Cooperative Test Division of the Educational Testing Service. The test was answered by the subjects on entering the study and again after a calendar year.

An assumption under which repeated measurements designs can be applied to item responses is that the variance-covariance matrices of the responses for the various populations are equal and symmetrical (2, 3, 4, 5, 7, 8, 10, 11). To ensure that the items furnishing the repeated measurements conformed as much as possible to this condition, the pretest responses were factor and cluster analyzed. The results (6) led to a split of the STEP-Math 2A into three parts containing 16, 19, and 15 items. The first part was highly homogeneous (KR-20 of .85). As such, its items were used as constituting the various levels of factor D. Since all the levels of each of the four factors were included in the experiment, the analyses were done in terms of a fixed model. The design was applied separately to the pre and the post responses.

RESULTS

The results of the analyses are given in Table 1. Fifteen rows of Table 1 are numbered, each representing a testable hypothesis. Instead of describing the results related to all the fifteen, an attempt is made here to deal with only those hypotheses in which the test items showed interactions with the classificatory factors and their implications for the 'score' hypotheses. They are:

(a) Item-teacher interaction was found both in the pretest and the posttest data (row no. 9). This indicates that when the experiment began, the students taught by the various teachers did not have uniform background in relation to the subject-matter areas covered by the test items and that the situation had not changed at the end of the experiment which was after a year. The finding has direct implication for the 'score' hypothesis on differences between the teachers (row no. 1). Because of the presence of item-teacher interaction, 'between teachers' differences should be approached through an examination of 'simple' main effects rather than that of the 'overall' main effect.

(b) There was item-sex interaction at the pretest stage (row no. 10). Further data analysis seems to show that the items were classifiable in three subgroups such that on the first group the boys did better, on the second equally well, and on the third worse, when compared to the girls. The observed absence of a sex difference (row no. 2) is thus misleading and reminds one of Moonan's example cited earlier.

Here again, an examination of 'simple' main effects is necessary.

(c) Item-text interaction was present on the pretest occasion (row no. 11) showing that when entering the investigation, the students of the SMSG text had better preparation than their counterparts on some items and that the reverse was probably true of other items. Item-text interaction may have arisen from the failure of about a third of the teachers to allocate the students randomly to the two types of texts. Its presence, however, does imply that

TABLE 1

PRE AND POST ANOVA TABLE FOR FOUR FACTOR DESIGN: REPETITIONS ON ONE FACTOR

			Pretest		Posttest	
	SV	df	MS	F	MS	F
(1)	A:Trs.	14	2. 216	27. 56**	3. 323	37. 17**
(2)	B:Sex	1	. 170	2. 11	. 011	. 00
(3)	C:Texts	1	. 002	. 03	. 002	. 02
(4)	AB:Tr. -Sex	14	. 050	. 62	. 045	. 51
(5)	AC:Tr. -Text	14	. 119	1. 49	. 191	2. 14**
(6)	BC:Sex-Text	1	. 012	. 15	. 009	. 10
(7)	ABC:Tr. -Sex-Text	14	. 052	. 65	. 042	. 47
	Sub. w. groups	453	. 0805		. 0894	
(8)	D:Items	15	8. 410	55. 40**	10. 119	71.56 **
(9)	AD:Tr. -Item	210	. 361	2. 38**	. 317	2. 24**
(10)	BD:Sex-Item	15	. 270	1. 78*	. 179	1. 27
(11)	CD:Text-Item	15	. 266	1. 75*	. 115	. 81
(12)	ABD:Tr. -Sex-Item	210	. 125	. 82	. 159	1. 12
(13)	ACD:Tr. -Text-Item	210	. 159	1. 05	. 126	. 89
(14)	BCD:Sex-Text-Item	15	. 136	. 89	. 030	. 21
(15)	ABCD:Tr. -Sex-Text-Item	210	. 172	1. 13	. 111	. 79
	DxSub. w. groups	6795	. 1518		. 1414	

*Significant at . 05.

**Significant at . 01.

the acceptance of the null hypothesis related to the differences between the texts (row no. 3) is not justifiable.

Since item-text interaction was not found in the posttest data, one could also conclude that the texts had something to do with eliminating the item-text interaction present in the pretest data. In this sense, then, the texts did differ, in spite of the conclusion to the contrary from row no. 3 of Table 1.

DISCUSSION

In all the researches in which performance on an objective achievement test furnishes the criterion measures likely to lead to certain conclusions, the researcher has two major choices: (i) He can use all the items of the test as a group and derive a single score for each subject. This is the most common practice. (ii) He can use each item independently and derive as many 'scores' or measures for each individual as there are items. (ii) is essentially an expension of (i) in the sense that an additional factor is added to the experimental design implicit in (i) and in the process, (i) is transformed into a repeated measurements design.

The extra computational labor involved in the application of a repeated measurements design is only a negligible fraction of the total cost of the research in the age of computers and can pay rich dividends. That this is so is clearly supported by the evidence presented here. From the conclusions cited above, it is clear that whenever test items show interaction with one or more of the "factors of experiment" (9), many of the conclusions derived through the use of alternative (i) would need to be re-examined through an analysis of 'simple' main effects. Alternative (ii) is, therefore, recommended for comparing 'treatments.'

FOOTNOTE

1. Grateful acknowledgements are due to the Minnesota National Laboratory, St. Paul, Minnesota, for supplying the data for the study and to the Alberta Advisory Committee on Educational Research for financing the analyses.

REFERENCES

1. Allen, F. B., and others, First Course in Algebra, Parts I and II, Yale University Press, New Haven, 1960.

2. Box, G. E. P., "Non-normality and Tests on Variance," Biometrika, 40:318-335, 1953.

3. Box, G. E. P., "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems. I. Effect of Inequality of Variance on the One-Way Classification," Annals of Mathematical Statistics, 25:290-302, 1954a.

4. Box, G. E. P., "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems. II. Effects of Inequality of Variance and of Correlation Between Errors in the Two-way Classification," Annals of Mathematical Statistics, 25:484-498, 1954b.

5. Cochran, W. G. and Cox, Gertrude, M., Experimental Designs, (2nd ed.), Wiley, New York, 1957.

6. Gupta, R. K., "Certain Techniques of Multivariate Analysis Applied to Different Measures of Inter Item Relationships for Developing Unifactor Tests," Journal of Educational Measurement, 5:223-230, 1968.

7. Kogan, L. S., "Analysis of Variance—Repeated Measurements," Psychological Bulletin, 45:131-143, 1958.

8. Kogan, L. S., "Variance Designs in Psychological Research," Psychological Bulletin, 50:1-40, 1953.

9. Moonan, W. J., "Simultaneous Examination and Method Analysis by Variance Algebra," Journal of Experimental Education, 23:253-257, 1955.

10. Nandi, K. H., "Mathematical Set-up Leading to Analysis of a Class of Designs," Sankhya, 8: 172-176, 1947.

11. Winer, B. J., "Statistical Principles in Experimental Design," McGraw-Hill, New York, 1962.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

A STUDY OF CLUSTER GROUPING

SUZANNE W. LARSEN
The University of Tennessee

JAMES R. MONTGOMERY
Virginia Polytechnic Institute

ABSTRACT

The increasing size of most universities in the United States at times has led to efforts to counteract largeness by creating many small units within the larger one. The following study reports on a pilot study to evaluate the effectiveness of a grouping pattern in which freshman students had their class schedules arranged in such a way as to have at least three sections (courses) in common with others in their respective group.

THE NEED for some new ideas concerning student group identification, such as cluster grouping, results at least in large measure from the tremendous growth of student enrollment in American institutions of higher education in the last quarter of a century (11). This steadily increasing enrollment in many institutions of higher education has been a concern to administrators who fear that increase in the size of institutions will be detrimental to the intellectual growth and personal development of the students (9). One of the proposed solutions to the problem has been clustering students into various types of relatively small groups such as groups based on common residence, common course sections, or both (1, 5, 11). Some of the objectives of these types of groupings have been: 1) better and faster adjustment to college by new students, 2) lowering the dropout rate, 3) closer professor-student relationships, and 4) more meaningful coursework as a result of coordinating assignments, for example, an English theme might be written on the history topic for that day.

History of Grouping Students

Features of cluster grouping have had a long history in the English universities of Oxford and Cambridge (11). The Claremont Colleges were among the first in the United States to attempt to have students share facilities yet retain their identity with a smaller group. The University of the Pacific first applied the term, "cluster college," to a group of undergraduate institutions on a single campus. The first campus to be planned and built around the cluster college idea was the Santa Cruz campus of The University of California (11). The cluster college idea basically provides that a given number of students live and usually receive instruction in a common group of facilities. But since many institutions have a large number of commuting students or have facilities not designed to accommodate small living-learning units, interest has at times developed in structuring cluster groups (as opposed to cluster colleges) which would still serve as primary groups within the context of the large university.

Related Research

Several universities throughout the United States currently are exploring the idea of cluster grouping or have already involved themselves in such a program (2, 3, 5, 9, 10, 11). However, the only objective evaluation of cluster grouping in the literature search appears to be one undertaken in The Florida State University (9). Researchers in Florida State report measures of extra curricular participation, informal social participation, time devoted to studying, perceived stress, and grade-point average showed no systematic differences between groups. While the experimental groups had a lower attrition rate, Stakenas (9) raised the question of a possible sampling error. General satisfaction data, for example, in terms of satisfaction with courses, dormitory social activities, new friendships, were higher for the clustered students ($p < .025$) at the end of the first term, but the difference did not remain significant ($p = .10$) at the close of the school year. Of particular interest was the decrease in satisfaction expressed by the living-learning clusters and the increase in satisfaction expressed by both the randomly-housed clusters and the nonclustered groups. Measures of perceived peer pressure were found to be greater in the clustered groups.

Problem

Recently the administration of The University of Tennessee (U. T.) became actively interested in the idea of grouping students. Student enrollment had been increasing rapidly and the dropout rate was of increasing concern. As a consequence a pilot study was developed for the Fall Quarter, 1967, to investigate the effects of cluster grouping.

Definition of Cluster Grouping

First-quarter freshmen enrolled in the College of Liberal Arts, having a common residence, and having three common course sections made up the cluster groups. First-quarter freshmen were chosen because one of the proposed and expected effects of cluster grouping was a better and faster adjustment to college. Using new freshmen eliminated the contamination of prior college experience. Students in the College of Liberal Arts, which has diverse course offerings even at the first-year level, comprised the experimental and control groups. Other colleges tend to have a natural course grouping for their students by virtue of the requirements within the colleges. A newly opened residence hall, with separate wings for men and women in a new dormitory complex, served as the place of residence for three reasons: better environmental control existed due to a board-and-room combination; it was assumed that students living in this area would be from a somewhat similar socio-economic background; a residential control group could be selected from two similar dormitories in the complex.

Students and their advisers selected the courses to be taken. Following this selection, schedules were developed so that each experimental group had at least three common courses with identical section assignments.

METHODOLOGY

Experimental Groups

Five experimental groups were formed based on three common courses. Freshman courses in English and History served as two of the three courses in each of the five groups. The third course for Group 1 was French, for Group 2 Chemistry, for Group 3 Zoology, for Group 4 elementary Spanish, and for Group 5 intermediate Spanish. Since the class size for English is restricted, a maximum size limit of 27 was set for each experimental group.

Final group membership. In order to be qualified for group membership, as previously stated, status, residence, and membership in three experimental sections had to be met by each student to be considered a member of the cluster group. While very few changed residence plans before Fall Quarter, many of the experimental members had to be dropped from the groups because of requested section changes or dropping one or more of the three scheduled cluster classes. No attempt could be made to encourage students to stay with their group, because one of the variables to be studied was the rapidity with which they became aware of their group membership. In addition, final group membership depended upon the return of a questionnaire used in the study.

Control Groups

Residence control (RC). Freshmen liberal arts students from two other residence halls in the complex composed the RC group. In order to keep the proportion of men and women equivalent to the experimental groups, every fifth woman and every third man was selected from an alphabetized listing of qualified subjects. Again, to remain a member of this group, a questionnaire had to be returned.

English control (EC). In order to consider the difference between student-instructor relationships in a cluster grouping study, it appeared of interest to introduce another factor. Of the five instructors in the experimental English sections, three had other freshman English sections. These sections were taught by the teachers who had experimental groups 1, 3, and 5. Reference will be made hereafter to EC1, EC3, and EC5 while the experimental English groups will be indentified by EE1, EE2, EE3, EE4, and EE5.

Other Students Surveyed

Additional information was obtained by means of the questionnaires from those students scheduled in the EE classes but not members of the cluster groups. In many cases these were former members who were dropped because of not fulfilling all the final requirements. Some, however, were never considered for membership. These groups will be referred to as "outsiders" and will be designated as OE1, OE2, OE3, OE4, and OE5.

Instrument Used

An instrument was designed especially for this study to attempt an assessment of each student's connotative meaning of the concepts, "My University," "My English Class," and "Myself," as measured by a semantic differential technique (6). Adjective pairs from those used by Pervin (7) were placed on a 7-point scale to compare the concepts, "My University" and "Myself," as well as "My English Class" and "Myself."

This instrument also collected such data as sex, age, size of high school graduating class, geographical location of high school, fraternity or sorority affiliation, and whether The University of Tennessee had been the student's first choice in applying to college. Students were asked about their awareness of being a member of a clustered group, and, if so, when they became conscious of such clustering. The number of courses taken during the quarter was requested as well as the number of visits home. The respondents were asked to express in percentages the amount of time spent in studying, recreation, socializing, working, and other. A final question asked them to check the one group in which most of their friends were located with the choices being classmates, dormmates, fraternity-sorority, high school friends also enrolled at The University of Tennessee, non-university people, and other.

The questionnaires were mailed to the members on November 17, 1967. Follow-up letters were sent a week later to those not having yet responded. Additional questionnaires were mailed to those requesting them after having received the follow-up letter and not able to find the original.

Analysis of Data

Semantic differential scales. The seven points on each scale were assigned the sequential numerical value 1 through 7. Instructions were to circle one number for each adjective pair or scale. The number circled was the score assigned to that scale.

Concept comparison scores. A distance analysis, as suggested by Osgood (4) was used to study the relation of two concepts. If they were alike in meaning for the individual, the distance in semantic space would be small and vice versa. Pervin (7) found similarity in meaning related to satisfaction.

The difference between the two scores on the same scale for two concepts for one individual was designated as "d." The linear distance, over all scales, between any two concepts, the "D" statistic (6), was found for each individual using the formula: $D = \sqrt{\Sigma d^2}$.

For each individual a D_1 score for the concept comparison, "My University" and "Myself," and a second score, D_2, representing the comparison between "My English Class" and "Myself" were calculated.

Group comparison scores. In order to determine the connotative distance between the two concepts as expressed by the group, D scores were averaged ($\overline{D}$) across individuals for each group. Then t-tests were used to analyze the $\overline{D}$ scores for the experimental and control groups. Grade point averages (GPA) for Fall and Winter Quarters were also used to compare the experimental and control groups, again using t-tests.

Ordered individual scores. The D_1 scores of all experimental and control members were ordered from the smallest or lowest (a similar meaning for the concepts, "My University" and "Myself") to the highest and total group was divided into thirds, termed hereafter the bottom (lowest) third, middle third, and highest third. Chi-square values were calculated to determine if there were any differences in personal data characteristics for the lower or higher third. Cells containing less than five members were collapsed before the chi-square values were determined.

RESULTS

Group Size

From an original N of 107, fifty-five of the experimental group were qualified for cluster group membership. Questionnaires were returned by 91 percent of those qualified, leaving a total of fifty experimental group members. These were divided into their respective groups as follows: EE1--8, EE2--15, EE3--13, EE4--7, EE5--7.

Of the seventy-one students selected for membership in the RC group, 73 percent or fifty-two students returned the questionnaire. Eighty-one percent (N = 22) of the females and 68 percent (N= 30) of the males responded.

TABLE 1

COMPARISON OF GROUPS USING SEMANTIC DIFFERENTIAL DATA

Groups	$\overline{D}_1$	SD	df	t	p
Residential Control					
EE1, EE2, EE3, EE4, EE5 (N=50)	8.5332	2.4456	100	0.0080	n.s.
RC (N = 52)	8.5294	2.3170			
Groups	$\overline{D}_2$	SD	df	t	p
English Control					
EE1 (N = 8)	5.2362	1.1176	23	-2.9867	n.s.
EC1 (N = 17)	4.0029	0.8871			
EE3 (N = 13)	6.3562	2.3596	27	-0.9133	n.s.
EC3 (N = 16)	5.5644	2.2909			
EE5 (N = 7)	3.7000	1.3039	23	1.5177	n.s.
EC5 (N = 18)	4.7917	1.7111			

Questionnaires were returned by 69 percent of the EC group members. The final group sizes were EC1--17, EC3--16, and EC5--18.

The number of non-cluster members or "outsiders" responding to the questionnaire was: OE1--14, OE2--4, OE3--7, OE4--13, OE5--12.

Group Concept Comparison Scores

No significant difference (t, $p > .05$) was found between the experimental group and the residential group using the semantic differential data (Table 1). For the EE3 and EC3 as well as the EE5 and EC5 groups no differences were found. For the expected direction (one-tailed t-test) no significant difference ($p > .05$) was found between the EE1 and EC1 groups. However, the absolute magnitude was such that had a two-tailed t-test been indicated, the control group would have had significantly ($p < .01$) closer concept meanings between "My English Class" and "Myself" (Table 1).

Using the D_2 score as the basis of comparison, no difference was found between the clustered and nonclustered members in the EE sections (Table 2).

GPA Comparison

In comparing the experimental and control groups on Fall and Winter GPA's, the only significant difference ($p < .05$) was again a reversed difference with EC1 having a higher GPA than EE1 in the Fall Quarter. This did not hold into the Winter Quarter, however (Table 3).

TABLE 2

COMPARISON OF SEMANTIC DIFFERENTIAL DATA FOR CLUSTER AND NONCLUSTER MEMBERS IN AN ENGLISH CLASSROOM

Groups	$\overline{D}_2$	SD	df	t	p
EE1 (N = 8)	5.2362	1.1176	20	0.3916	n. s.
OE1 (N = 14)	5.6300	2.6918			
EE2 (N =15)	4.3533	1.9019	17	-0.2561	n. s.
OE2 (N = 4)	4.0900	1.4256			
EE3 (N = 13)	6.3562	2.3596	18	-0.0311	n. s.
OE3 (N = 7)	6.3200	2.6947			
EE4 (N = 7)	5.5886	1.7475	18	-0.7889	n. s.
OE4 (N = 13)	4.7562	2.4641			
EE5 (N = 7)	3.7000	1.3039	17	1.2580	n. s.
OE5 (N = 12)	4.7175	1.8821			

Personal Information Data Characteristics

Of the fifteen personal information data variables considered, only two characteristics were found to have significance (Table 4). Sex was significant ($\chi 2$, $p < .05$) with a greater percent of men in the bottom third of the ordered D_1 scores (Table 5). The amount of time spent studying was the second significant variable ($\chi 2$, $p < .025$). Percentages for these data are reported in Table 6.

DISCUSSION AND SUMMARY

If the cluster grouping of students had given its members an identity group, an attitude of greater warmth toward or satisfaction with the university would have been expected, as expressed by a difference between the EE groups and the RC group on $\overline{D}_1$ scores. A closer student-instructor relationship as a result of cluster group membership would have been reflected by the EE groups having a closer meaning for the two concepts, "My English Class" and "Myself," than the comparable EC groups.

These differences did not emerge. Therefore, we must assume that cluster grouping as measured in this study proved ineffective. However, it should be pointed out that this experimental survey had certain weaknesses in the design: 1) A disclosure to the experimental group members of their participation in a cluster grouping study was made after the questionnaires were mailed but before all were returned. Such a disclosure should have increased the probability of getting significant results as it has been found (8) that increased attention to a group subsequently results in increased satisfaction for the group. 2) The liberal drop-add and section-change policy of the university allowed the cluster groups to decrease to such size that they were the minority group even in the small English classes, thus destroying some of the possible effects of the treatment. 3) Three of the courses had extremely high drop-out rates and should not have been included in the group course clusters. 4) The control groups were not matched for these troublesome courses.

The only two significant findings for group comparison were reversed with the control group exceeding the experimental group. These involved the same groups, EC1 and EE1. Since the differences involved GPA as well as satisfaction, assuming satisfaction to be measured by similar concept meanings, an unmeasured variable, intelligence, may be interacting. Perhaps the more intelligent students adjust faster and better to the college environment.

TABLE 3

GPA COMPARISONS FOR EXPERIMENTAL AND CONTROL GROUPS FALL AND WINTER QUARTER

Groups	Mean GPA	SD	df	t	p
Residential--Fall					
EE1, EE2, EE3, EE4, EE5	2.0000	0.7317	97	-1.5562	n. s.
RC	2.2188	0.6664			
Residential--Winter					
EE1, EE2, EE3, EE4, EE5	2.2098	0.6655	97	1.3690	n. s.
RC	2.0038	0.8219			
English--Fall					
EE1	1.5237	0.4158	22	-2.2556	n. s.
EC1	2.1244	0.6885			
EE3	2.0517	0.7839	25	0.2068	n. s.
EC3	1.9887	0.7889			
EE5	2.5600	0.6686	22	1.5982	n. s.
EC5	2.0194	0.7826			
English--Winter					
EE1	1.7762	0.7509	22	-1.3495	n. s.
EC1	2.2031	0.7208			
EE3	2.1567	0.6437	25	0.9439	n. s.
EC3	1.8380	1.0156			
EE5	2.5914	0.6575	22	1.0109	n. s.
EC5	2.1918	0.9505			

The nonclustered members in the experimental English sections did not demonstrate greater dissatisfaction by dissimilar concept meanings; therefore, we will assume they did not feel left out of "the group." In three of the five groups, however, they were twice as large as the experimental group, and, therefore, would have no reason to feel like outsiders.

No difference between the experimental and control groups on GPA supported the results found by Stakenas (9).

No explanation for the higher percentage of men than women in the more satisfied group can be suggested from the results here. The fact that those more satisfied with their college life spend more time studying is an interesting fact that might be worthy of future research.

From the findings reported in this study as well as those reported by Stakenas (9), future consideration of the use of a cluster grouping technique would seem to be of little value as a means of providing better and faster student adjustment to college. If, however, further study of cluster grouping were desired, these suggestions appear relevant: 1) Proximity within the residential areas might prove more effective than grouping around English. 2) Membership in a clustered group should be put on a voluntary basis in order to maintain the groups. 3) A careful evaluation of courses should be made before they are included in the clustered group. If at all possible, an attempt should be made by the instructors of the clustered courses to coordinate assigned material as a means of enriching the overall coursework for the clustered groups. Although this would involve considerably more time and effort on the part of the professors involved, this enrichment might entice more students to become members. It is possible, however, that the enriched courses would attract only the brighter students, not a representative group.

TABLE 4

PERSONAL INFORMATION DATA FOR ORDERED INDIVIDUAL SCORES DIVIDED INTO THIRDS

Variable	df	χ^2
1. Sex	2	6.2072*
2. Size of high school	4	5.2706
3. Location of high school	4	0.2241
4. Fraternity-sorority affiliation	2	0.2065
5. U. T. as first-choice college	2	2.2969
6. Knowledge of group membership	2	0.6798
7. Number of courses in schedule	2	2.5342
8. Residence (home vs. other)	-	0.0000[a]
9. Number of visits home	4	2.2744
10. Amount of time spent studying	10	20.9559**
11. Amount of time spent in recreation	6	1.9348
12. Amount of time spent socially	6	3.1186
13. Amount of time spent working	0	0.0000[a]
14. Amount of time spent otherwise	1	0.0168
15. Group in which friends are listed	4	3.9176

[a] Insufficient data.
*Significant at .05 level.
**Significant at .025 level.

REFERENCES

1. Crawford, E. M., "Universities Grow Large, but Keep Education Student-sized," College and University Business, 41:41-43, (6), 1966.

2. Grant, C. W., "An Experiment in Clustering," unpublished research proposal, University of Utah, 1967.

3. Jarrett, J. L., "The Small College on the Big Campus: Santa Cruz After One Year," Saturday Review, 1967, January 21, 67-79.

4. Kerlinger, F. N., Foundations of Behavioral Research, Holt, Rinehart, and Winston, New York, 1967, chapter 32.

5. Mayhew, L. B., "The New Colleges," in S. Baskin (Ed.), Higher Education: Some Newer Developments, McGraw-Hill, New York, 1965, pp. 1-26.

6. Osgood, C. E., Suci, G. J., and Tannenbaum, P. H., The Measurement of Meaning, University of Illinois Press, Urbana, 1957.

7. Pervin, L. A. "A Twenty-college Study of Student x College Interaction Using TAPE (Transactional Analysis of Personality and Environment), The Journal of Educational Psychology, 58:290-302, (5), 1967.

8. Roethlisberger, F. J., and Dickson, W. J., Management and the Worker, Harvard University Press, Cambridge, 1939.

9. Stakenas, R., "The Effect of Correlated Class Scheduling and Dormitory Housing on the Social and Academic Behavior of College Freshmen," paper presented at the 52nd Annual Meeting of the American Educational Research Association, February 1968, Chicago, Illinois.

10. Time, "Living-learning Clusters," 1966, September 9.

11. Woodring, P., "The Small College on the Big Campus: The Idea of Cluster Colleges," Saturday Review, 1967, January 21, 64, 66, 81.

TABLE 5

PERCENTAGE OF MEN AND WOMEN IN EACH THIRD OF ORDERED INDIVIDUAL SCORES

Ordered Scores	Men	Women
Top third	29.2	40.4
Middle third	30.2	38.6
Bottom third	40.6	21.1
Total	100.0	100.0

TABLE 6

PERCENTAGE OF TIME SPENT STUDYING IN EACH THIRD OF ORDERED INDIVIDUAL SCORES

	Percent of Time Spent Studying					
Ordered Scores	10-20	30	40	50	60	70-90
Top third	69.6	30.8	30.4	37.5	17.6	24.2
Middle third	17.4	34.6	34.8	33.3	58.8	30.3
Bottom Third	13.0	34.6	34.8	29.2	23.5	45.5
Total	100.0	100.0	100.0	100.0	100.0	100.0

ders

DEMBAR
EDUCATIONAL
RESEARCH
SERVICES, INC.

POST OFFICE BOX 1148 · MADISON, WISCONSIN 53701

PERTINENT RECENT TITLES FROM DERS

THEATER IN AMERICA

By Prof. Robert E. Gard, Prof. Marston Balch, and Pauline Temkin
212 pages $6.50 hardcover, $4.95 softcover
The most complete story of contemporary theater published to date. Commissioned by the National Theater Conference.

FUN WITH GAMES

By Prof. John E. Anderson
188 pages $4.95
If a suitable recreational activity isn't described and explained in this book, it probably doesn't exist.

WISCONSIN SIDEROADS TO SOMEWHERE

By Prof. Clay Schoenfeld
266 pages $2.95
A collection of stories and essays in the spirit of Thoreau, Muir, and Leopold. Offers entertaining, informative reading about the adventures in outdoor recreation and conservation that are everybody's for the finding down the sideroads of America.

RESEARCH AND DEVELOPMENT TOWARD THE IMPROVEMENT OF EDUCATION

Edited by Profs. Herbert J. Klausmeier and George T. O'Hearn
176 pages $5.75 softcover
A compilation of papers by 21 leading educators devoted to imaginative strategies for cognitive-learning research and the resulting instructional practices.

IDEAS AND IMAGES

By Prof. Lindley J. Stiles
96 pages $3.50
Popular poetry about and for young people—and for all who share with youth the exciting venture of growing up.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

A STUDY OF SUPERVISORS' RATINGS OF MOST EFFECTIVE AND LEAST EFFECTIVE INDUSTRIAL ARTS TEACHERS ON THREE COMPETENCY DIMENSIONS[1]

MORTON MARGULES and VIRGINIA M. KEITH
University of Ottawa

ABSTRACT

The problem was to determine the relationship between industrial arts supervisors' judgments of teacher effectiveness and their judgments of teacher competency on the dimensions of interpersonal relations, teaching techniques, and technical proficiency. A device was developed to obtain the supervisors' perceptions of their most and least effective teachers. The data obtained were used to determine the relationship of the judgments of effectiveness to scores on the dimensions. Analysis of variance and Scheffe' S-method were used to analyze the data. The conclusions were: teachers are judged more competent in technical proficiency; the least effective teacher has greater deficiencies in interpersonal relations and teaching techniques; and the teaching techniques dimension differentiates, to the greatest degree, the most from the least effective teacher.

RESEARCH ON teacher effectiveness has frequently made use of supervisors as raters. Yet, after ratings have been obtained for use in the study, the supervisors making the ratings have been forgotten. While teacher effectiveness has been the subject of multitudinous studies by Barr (1), Ryans (9), and others, a major facet lacking consideration has been the supervisor's approach to teacher evaluation.

This investigator attempted to relate teacher effectiveness, as viewed by the supervisor, to dimensions of the teaching process which were derived from an analysis of the instructional process by observation and discussions with leaders in the field. Recognizing that the nature of the subject matter occasions differences in the teaching process, it was deemed advisable to consider teaching in a single field. Therefore, the particular field of industrial arts was selected.

Industrial arts is considered a unique area of education because behavioral changes in students are sought by the use of course content and facilities which closely relate to our highly industrialized society. In view of the emphasis being placed upon this area of education, it was appropriate to seek information concerning the approach of the industrial arts supervisor to evaluating teachers. The studies of Ehrenborg (3), Schank (11), and others, showed that supervisors use a variety of factors to assess teacher effectiveness. Yet it would seem reasonable that the industrial arts teacher's ability to relate to others, his ability to use a variety of teaching techniques, and his grasp of appropriate technical concepts and skills are dimensions which should determine his effectiveness. Therefore, the purpose of this study was to determine the degree to which industrial arts supervisors consider such dimensions as interpersonal relations, teaching techniques and technical proficiency, in their evaluation of industrial arts teachers and how these dimensions relate to judgments of teacher effectiveness.

The scope of this study is limited by the following: (1) the approach to the complex problem of industrial arts teacher effectiveness is made in terms of the three dimensions stated; (2) it was not the purpose of this study to determine which of these dimensions is most needed for effective teaching. Thus, this was not a study of the behavior of teachers but a study of the nature of supervisors' perceptions of and judgments about the behavior of teachers.

Significant contributions of this study may be:

1. to show a relationship between judgments of effectiveness and underlying generic dimensions.
2. to encourage others to make comparisons across subject matter areas of the various dimensions to teacher effectiveness.
3. to provide a basis for training supervisors to make objective judgments of teacher effectiveness.
4. to provide a basis for the development of instruments upon which supervisors can render objective judgments of teacher effectiveness.
5. to stimulate educators of industrial arts teachers to consider supervisor's expectations of teacher behavior.

RELATED RESEARCH

A search of the literature revealed no study which sought the global teaching dimensions which may be used by a supervisor when he makes a judgment about teacher effectiveness. In the training of industrial arts teachers, desirable teaching techniques and technical proficiency are an accepted part of the curriculum. Interpersonal relations, as such, are not formally included in the teacher training program. Thus, it would seem that more weight would be given to technical proficiency and teaching techniques than to interpersonal relations when the supervisor is considering teacher effectiveness.

With the exception of technical proficiency, the other two dimensions stated above have been considered by most studies. But technical proficiency is a most important ingredient when we consider industrial arts teacher effectiveness. It is even more important when we consider its significance in the supervisor's criteria for evaluating the teacher. No study has been undertaken which specifically uses the dimensions this investigator has devised to determine what considerations might apply when the supervisor of industrial arts evaluates teachers in his charge.

HYPOTHESES

The null hypotheses investigated in this study were:

1. When the three dimensions of interpersonal relations, teaching techniques, and technical proficiency are considered cummulatively, teachers rated most effective will not receive significantly higher overall ratings of competency than teachers rated least effective.
2. All three dimensions will prove equally significant in discriminating between most effective and least effective industrial arts teachers.

METHODOLOGY

Sample

The sample consisted of randomly selected industrial arts supervisors who had a minimum of 5 years experience in this capacity and who, by virtue of their assigned location, were in daily contact with their supervised teachers.

A questionnaire sent to the supervisors provided an unstructured opportunity for them to indicate those characteristics which they felt the ideal industrial arts teacher should possess. Answers were guided only to the extent that they were categorized within the three dimensions of the inquiry.

Preparing the Dimension Scales

The three dimensions used were interpersonal relations, teaching techniques, and technical proficiency. The operational definitions follow:

1. Interpersonal relations refers to everything that "goes on" between one person and another (or others) by way of perception, evaluation, understanding, and mode of reaction.
2. Teaching techniques refers to the processes involved in the action necessary to establish a common bond between teacher and learner.
3. Technical proficiency refers to the knowledge, skill, and degree of expertness with which the principles of industrial arts, mechanical arts, and sciences are applied.

The following steps were taken to develop and validate the observation instrument:

- Statements were gathered from the literature and the responses of the supervisors to the questionnaire. Abstract, ambiguous, irrelevant, and blanket items were removed and the statements made objective, workable and behavior-oriented.
- The remaining items were then submitted to a jury. This group of judges was requested to eliminate those items which they felt had no value in determining industrial arts teacher effectiveness in each of the three dimensions. They were then asked to rank order the remaining items on the basis of their value for ascertaining the most effective industrial arts teacher.
- The ranked frequencies were then established. The mean for each item was calculated and all items were placed in a final ranking. A total of 109 items were derived--thirty-seven in both the interpersonal relations and technical proficiency dimensions and thirty-five in the teaching techniques dimensions.
- The items were phrased so that half of the favorable responses were in positive terms, and the other half of the favorable responses were in negative terms. It was expected that this procedure would discourage the development of a response set on the part of the supervisors. The items were placed in a Likert-type scale of one to five. The value of five was used to indicate the most effective category. The items were then randomly ordered in the pilot device. A random selection of ten supervisors were then asked to rate their most effective and least effective teacher. Effectiveness scores on twenty teachers were thus obtained.
- An item analysis was conducted according to the Flanagan method (4) to determine the discriminatory power of each item and to establish the internal validity of the device.
- After the item analysis was completed, a total of sixteen usable items for each dimension were selected. Flanagan correlations ranged from .70 to .87. The final Schedule, was then developed by using these items.
- The final instrument was given to each of the original ten supervisors with a request to rerate the same teachers. Four weeks elapsed between the use of the pilot instrument and the final instrument. The items of the second testing were then correlated with the corresponding items on the original scale using the Flanagan method.
- Based upon the dual criteria of consistency on the test-retest and the Flanagan r, two items in each

dimension scale were dropped due to low r's. Correlations on the forty-two remaining items ranged from .56 to .87.

The instrument was then sent to fifty-seven randomly selected supervisors who met the selection criteria. Fifty-two were usable for the study and provided a 91 percent return. The instrument provided the raw scores for 104 teachers on the three competency dimensions. A total raw score for each teacher was obtained by adding the scores given by the supervisor on each of the competency dimensions.

In order to assess rater validity, previous administrators ratings of the participating teachers were correlated with the scores of overall effectiveness obtained from the supervisors.

Analysis Plan

A 2 x 3 analysis of variance with repeated measures on the factor was used to analyze the data.

Three effects were tested:

1. the main effect of effectiveness rating (A)
2. the main effect of the specific competency area (B)
3. the interaction between the effectiveness rating and the specific competency area (AB)

For each significant effect the Scheffe' S-Method of comparing means was used.

The administrators' ratings were correlated with the scores on the three dimensions individually for the effective and ineffective teachers separately. This was done to compare the discriminability of the administrators' ratings with that of the supervisors.

RESULTS AND DISCUSSION

Table 1 provides the range and means of selected characteristics of the supervisors who participated in the study. If the supervisors in this study differ at all from the ordinary supervisor, it is in the direction of being more experienced, and perhaps, therefore, more sensitive. The Table also suggests that the supervisors, as a group, have been somewhat prepared for their role as upgraders of instruction which, per force, includes evaluation.

The results of the analysis of variance of effectiveness and competency dimensions appear in Table 2.

TABLE 2

ANALYSIS OF VARIANCE OF EFFECTIVENESS (A) AND COMPETENCY (B).

Source of Variation	SS	df	MS	F
Between subjects	51,736	103		
A (effectiveness)	26,017.66	1	26,017.66	103.18
Subjects within groups	25,718.34	102	252.14	
Within subjects	1,915	208		
B (competency)	703.82	2	351.91	74.01
AB	213.48	2	106.74	21.80
B x subjects within groups	997.70	204	4.89	

B. J. Winer, Statistical Principles in Experimental Design, New York, McGraw-Hill, 1962, p. 307.

Main Effect of Effectiveness

The main effect of the effectiveness rating was significant ($F = 103.18$, $p < .001$). An examination of Table 3 shows that the most effective teacher group had overall higher ratings across the three dimensions than the least effective teacher group. This was expected in view of the fact that the supervisors had *a priori* made the effectiveness selection and, therefore, were likely on any evaluative item to indicate their effective teachers were superior to their ineffective teachers.

TABLE 1

RANGE AND MEANS OF SELECTED CHARACTERISTICS OF SUPERVISORS (N = 52).

			Years of Professional Experience		Supervision	
	Age	Education[a]	Teaching	Supervision	Time Available[b]	No. of Teachers
Range	30-66	2-6	3-36	4-30	0-8	2-19
Mean	47.42	5.06	14.35	7.92	3.21	6.35

Supervisors' Inventory Form

[a] Education - 4 = Bachelor degree
5 = Master's degree or its equivalent
6 = Sixth year level equivalent

[b] Time available for supervision during the day.

TABLE 3

SUMMARY TABLE FOR TEACHER EFFECTIVENESS (A) AND TEACHER COMPETENCIES (B) MEAN SCORES (n = 52).

Effectiveness Rating	Competencies			
	Interpersonal Relations b_1	Teaching Techniques b_2	Technical Proficiency b_3	Sum
Most Effective a_1	62.6[a]	62.3	64.0	63.0
Least Effective a_2	44.6	41.9	47.6	44.7
Sum	53.6	52.1	55.8	

Note: All cell means differ significantly from one another (except between cells a_1 b_1 and a_1 b_2 and between a_1 b_1 and a_1 b_3) by the Scheffe' S-method.[a] Except for between cells a_1 b_2 and a_1 b_3 which differ at the .05 level, all other significant differences attain the .01 level of confidence.

[a] George A. Ferguson, *Statistical Analysis in Psychology and Education*, New York, McGraw-Hill, 1966, p. 296.

The great difference in means, however, suggests that there is a wide differentiation being made between these two effectiveness groups of teachers by the supervisor when he evaluates their behaviors in terms of an overall effectiveness score. Therefore, industrial arts supervisors do discriminate significantly between the most effective teacher and the least effective teacher when the three competency dimensions of interpersonal relations, teaching techniques, and technical proficiency are considered as a composite. On this basis, therefore, null hypothesis one was rejected at the .01 level of confidence.

Main Effect of Competency Dimensions

The main effect of the specific competency scores (B) was significant (F = 74.01, $p < .001$). It was shown that supervisors of industrial arts considered both the most effective and the least effective teacher groups as being significantly more competent in technical proficiency than in either interpersonal relations or teaching techniques. This was somewhat expected when the curriculum of industrial arts teacher education institutes is considered. Approximately one-third of the courses given are devoted to providing experiences designed to lead to technical competencies in the subject matter areas which comprise industrial arts education.

It was unexpected that among the three competency dimensions teaching techniques would be significantly lower than technical proficiency in the most effective teacher group and significantly lower than technical proficiency and interpersonal relations in the least effective teacher group. This suggests that our teacher education programs, although stressing the methodology of teaching industrial arts subjects, apparently are not adequately designed to provide the teaching techniques which supervisors deem necessary for their industrial arts teachers. Or perhaps the instructional methodology provided at the college is more in tune with contemporary thinking, and the supervisors are not familiar with the latest methods of lesson presentation.

It is also suggested, from a perusal of the overall means on the competency dimensions, that industrial arts teachers as a group are seen by their supervisors as relatively more effective in dealing with others than in presenting instructional material. Subtly, it may suggest that industrial arts teachers as a group are less articulate in presenting instructional material where demonstration and verbal communication are factors.

It was anticipated that among the three competency dimensions interpersonal relations would be significantly lower than the other two for each effectiveness group. This assumption was based upon the lack of formal curricula to develop this type of competency in the teacher education setting. Also, it was felt that industrial arts teachers as a group might be looked upon as more oriented toward objects and/or the task and less oriented toward people. This was clearly not the case for the least effective teachers who were seen as lowest on teaching techniques, while, for the most effective teachers interpersonal relations and teaching techniques were equally low. From a study of Table 3, supervisors do perceive their teachers as being more competent in technical ability than in either interpersonal relations or teaching techniques.

All three competency dimension means differed significantly at the .001 level of confidence.

Interaction Between Effectiveness and Competency Dimension

The interaction between the effectiveness rating and the specific competency dimension (AB) was significant ($F = 21.8$, $p < .001$). The data presented show that the most effective teacher group was far superior to the least effective teacher group on the three dimensions. The mean differences between the two effectiveness groups were extremely large for each dimension. However, relatively small differences were shown between competency scores within each group. In fact, between cell differences within each effectiveness group were numerically only 13 percent as large as mean differences between effectiveness groups. It appears that the sensitivity of the supervisor is small within effectiveness groups, but extremely large between the groups. The data suggests that it might be appropriate to repeat the study using a full range of teachers instead of two extreme effectiveness groups. This method might show less gross differences between effectiveness groups and greater differences within effectiveness groups on the three competency dimensions.

In judging the most effective teacher, only the competency dimension of technical proficiency was differentiated. A difference between interpersonal relations and teaching techniques was not obtained. This finding suggests that competency in interpersonal relations may be closely related to competency in teaching techniques when the most effective teacher is considered. Competency in teaching techniques may be related to the disirable interpersonal relations exhibited by the teacher. This may then lead to the point of view that highly desirable interpersonal relations are necessary to present instructional material effectively. In addition to this point of view, it can be inferred that the supervisor cannot tell the difference between interpersonal relations and teaching techniques when the most effective teacher is observed in action. It may be conjectured that in evaluating the most effective teacher, the supervisor merges all dimensions into the overall effectiveness rating, whereas in evaluating the least effective teacher, he uses more care in trying to determine the reasons for ineffectiveness.

The lowest mean for the least effective teacher in teaching techniques was quite unexpected. The additional courses in teaching methods required of industrial arts teachers purport to develop an understanding of the principles of teaching so that a variety of techniques, methods, and approaches may be adapted to effective teaching. One might conclude that what is gained from methods courses is not commensurate with the time spent on them. However, this conclusion should be tempered, because we are discussing the less effective teachers. The same results were not obtained with the more effective teachers.

Differences among the three dimensions were evidenced when the supervisors judged their least effective teachers. Least effective teachers tend to be rated as less competent in teaching techniques than in interpersonal relations and technical proficiency. One explanation may be that industrial arts teaching techniques are more easily observed and classified than interpersonal relations. For the ineffective teacher who is ineffective in all areas, it may be easier for the supervisor to perceive the more visible dimension of teaching techniques. It may also be that a modicum of technical proficiency cannot overcome deficiencies in teaching techniques.

It is apparent from the data that technical proficiency is not an adequate basis for considering a teacher effective in industrial arts. It is probably true that when evaluating teachers, supervisors perceive the more tangible outcomes of technical proficiency quite easily. Their perception may be enhanced by the fact that industrial arts teachers are, rightly or wrongly, considered to be technicians with highly developed technical abilities and skills. However, since the content and concepts of industrial arts are becoming more complex and involved due to the contemporary technology, supervisors may be viewing the teaching of industrial arts in new terms. They may now be looking for competencies other than those strictly manipulative acts and immediately discernible related knowledges. They feel rather that these other competencies should be part of the professional teacher's repertoire. The teacher needs to understand the basic principles of teaching and learning and he must be able to organize learning experiences by using a variety of methods appropriate to the lesson at hand.

It can be concluded that the most effective industrial arts teacher has exceptional competencies in the areas of interpersonal relations, teaching techniques, and technical proficiency. But of more concern, it can be concluded that the least effective industrial arts teacher has deficiencies in all three dimensions, but is particularly deficient in teaching techniques. It was shown that supervisors' scores for the least effective teachers featured greater discriminability among the three competency dimensions than is true for their scores for the most effective teachers.

Relationship of Supervisors' and Administrators' Judgments

A marked significant correlation, Table 4, was obtained between the supervisors' total scores and the administrators' ratings. However, a consideration of the correlations for each dimension for each of the effectiveness groups separately did not yield totally valid correlations. This was attributable to the use of only one-half of the total effectiveness data in computing such correlations. Therefore, the correlation within each cell was spuriously low. However, one can interpret the relative sizes of such correlations.

From these correlations, it may be concluded that supervisors and administrators are most inclined to agree on the dimension of teaching techniques as concerns the least effective teachers. Thence, more evidence for the visibility of this dimension among poor teachers. Of note also is the fact that agreement within cells is highest for the most effective teacher on technical proficiency, while for the least effective teacher agreement is highest on teaching techniques. Thus, a teacher is probably judged most effective because of his high technical proficiency.

It is also suggested from this analysis that industrial arts supervisors and administrators, although, significantly in agreement on their overall evaluation of industrial arts teachers, may be perceiving a different set of behaviors exhibited by the teachers when

TABLE 4

PEARSON PRODUCT-MOMENT CORRELATIONS BETWEEN SUPERVISORS' COMPETENCY SCORES AND ADMINISTRATORS' RATINGS CLASSIFIED BY SUPERVISOR'S EFFECTIVENESS RATINGS (n = 52)

Effectiveness Rating	Competencies			Total Score
	Interpersonal Relations b_1	Teaching Techniques b_2	Technical Proficiency b_3	
Most Effective a_1	.293	.383	.505	.421
Least Effective a_2	.333	.499	.458	.493
Total Score[a]	.635	.734	.663	.733

Lawrence T. Dayhaw, The Dayhaw Correlation Chart, Ottawa, Canada, University of Ottawa Press, 1956.

a N = 104.

Note: All correlations are significant at the .01 level of confidence except a_1 b_1.

these three dimensions are considered. Or they may be reacting to them differently. It further suggests that the technical proficiency dimension has relatively high visibility for both the supervisor and the administrator, particularly when compared to interpersonal relations. This may be due to the concrete nature of technical skills and knowledges. When considering interpersonal behaviors and teaching techniques, most administrators bring to bear an academic background and, consequently, an orientation somewhat different from the industrial arts trained supervisor. In spite of this difference in background, however, one is struck by the agreement rather than disagreement between judges. All correlations were significant at the .05 level of confidence.

CONCLUSIONS

This study has attempted to investigate the role of three competency dimensions in supervisors' assessment of industrial arts teacher effectiveness. It was not the intent of this study to delineate very specific and narrow characteristics, but to provide three broad underlying generic dimensions that are commonly recognized as competency areas in the teaching process. The results seem to lead to the following conclusions:

1. A study of supervisory practices in evaluating teachers is feasible.

2. Positive relationships exist between certain generic teaching dimensions and supervisors' judgments of teacher effectiveness.

3. The three competency dimensions, interpersonal relations, teaching techniques, and technical proficiency, do differentiate between the most effective and the least effective industrial arts teachers.

4. The least effective teachers have deficiencies on all three dimensions, but greater deficiencies in both interpersonal relations and teaching techniques.

5. The three dimensions do not equally differentiate between most effective and least effective industrial arts teachers.

6. Effective industrial arts teaching requires more than technical proficiency.

7. The teaching techniques dimension to the greatest degree differentiates the most effective from the least effective teachers.

8. Supervisors and administrators are probably influenced most by high technical proficiency when rating an effective teacher. On the other hand, they are primarily influenced by interpersonal relations and teaching techniques when determining the less effective teacher.

9. This approach to understanding the basis on which the supervisor evaluates his teachers and the considerations upon which his judgments are founded does not necessarily exclude the possibility that other dimensions are relevant and important. However, both the data and logic of the study support the assumption that the three dimensions studied are influential factors in ratings of effectiveness.

10. Expectancies of supervisors and administrators must be considered in teacher evaluation. While they tend to agree, they do not show complete overlap.

The findings of this study may also:

1. Reflect a need for teacher improvement programs in teaching techniques and interpersonal relations.

2. Provide a basis for training supervisors to make objective judgments of teacher effectiveness.

3. Provide a basis for the further development of instruments upon which supervisors can render objective judgments of teacher effectiveness.

4. Stimulate industrial arts teacher educators to consider supervisors' expectations of teacher behavior.

5. Make feasible a comparison across subject-matter areas of the relative contributions of the various dimensions to teacher effectiveness.

RECOMMENDATIONS

On the basis of the findings of this study it seems appropriate to make several recommendations:

1. Since very little research has focused on the problem of supervisory evaluation in industrial arts, it is suggested that within the basic framework of evaluation, further exploration be undertaken using the dimensional approach. The approach might also be explored in other subject-matter areas.

2. The device used in this investigation should be reviewed and refined.

3. Any new study undertaken should include a full range of effectiveness measures, so as to provide more refined data.

4. The recommedation made within and without education that the role of teaching techniques and methods be reduced received little support in this study. A serious mistake, at least with regard to industrial arts teaching, would be made in relaxing instruction in the teaching techniques area. Therefore, it is recommended that more attention be given to those courses which provide a vital understanding of the processes of learning and the methods of presenting industrial arts instructional material.

5. Exploration should be undertaken to seek ways of formally structuring into the teacher education curriculum programs which are intended to enhance positive interpersonal relations.

6. Finally, it is recommended that graduate level courses in evaluation be developed for present and future supervisors of industrial arts. These courses should have as their prime purpose the presentation of the latest information concerning subjective and objective teacher evaluation. This is particularly vital at this time in view of consideration being given to a shift from the guaranteed yearly increment to merit salary increases; objective evaluation must be on a firm and fair basis, easily understood by both the teacher and the evaluator.

FOOTNOTE

1. This manuscript is based on the writer's doctoral dissertation completed under the direction of Professor Virginia Keith, University of Ottawa, Canada, 1968.

REFERENCES

1. Barr, Alvin S., "The Measurement and Prediction of Teaching Efficiency: A Summary of Investigations," Journal of Experimental Education, 16:203-283, June 1948.

2. Brookover, Wilbur B., "The Relation of Social Factors to Teaching Ability," Journal of Experimental Education, 13:191-205, June 1945.

3. Ehrenborg, David, The Validation of a Device for Identification of Industrial Arts Teaching Success, unpublished doctoral dissertation, Los Angeles, University of California, 1963,xii-307 p.

4. Flanagan, John C., "General Considerations in the Selection of Test Items and Short Method of Estimating the Product-Moment Coefficient at the Tails of the Distribution," The Journal of Educational Psychology, 30:674-680, December 1939.

5. Giachino, J. W., An Analysis of the Success Qualities That Should be Emphasized in the Training of Candidates to Become Competent Teachers of Industrial Arts, unpublished doctoral dissertation, University Park, Pennsylvania State University, 1949, xi-205 p.

6. Medley, Donald M., and Mitzel, Harold E., "A Tentative Framework for the Study of Effective Teacher-Behavior," Journal of Experimental Education, 30:317-320, 1962.

7. Monroe, H. B., Personality Variables of Industrial Arts Teachers, unpublished dissertation, Greeley, Colorado, Colorado State College, 1960, xiii-78 p.

8. Morgan, J. B., Personality Variables of Industrial Arts Teachers, unpublished doctoral dissertation, Greeley, Colorado, Colorado State College, 1962, xi-135 p.

9. Ryans, David G., "Assessment of Teacher Behavior and Instruction," Review of Educational Research, 33:415-441, October 1963.

10. Ryans, David G., "Theory Development and the Study of Teacher Behavior," The Journal of Educational Psychology, 47:462-475, December 1956.

11. Schank, Kenneth Lloyd, Industrial Arts Supervision: A Synthesis of Selected Supervisory Principles and Practices, unpublished doctoral dissertation, College Park, University of Maryland, 1965, vi-266 p.

12. Selvidge, Robert W., "Success Factors in Teaching," The Industrial Education Magazine, 33: 305-306, June 1932.

13. Singer, Jr., Arthur, "Social Competence and Success in Teaching," Journal of Experimental Education, 23:99-131, December 1954.

14. Solomon, Daniel, "Teacher Behaviors Dimensions, Course Characteristics, and Student Evaluations of Teachers," American Educational Research Journal, 3:35-47, January 1966.

15. Torrence, Andrew Pumphrey, "A Study of the Relationships of Certain Competencies to Success in Teaching Vocational Agriculture," Journal of Experimental Education, 25:1-31, September 1956.

New Title From DERS

RESEARCH AND DEVELOPMENT TOWARD THE IMPROVEMENT OF EDUCATION

Edited by Herbert J. Klausmeier and George T. O'Hearn. 1968

176 pages **$5.75 softcover**

In this book 20 eminent scholars, researchers, and research administrators present their latest thinking about the processes and conditions of human learning, the processes and programs of instruction, and the application of research and devlopment strategies to educational problems.

As the editors point out in the Preface, the application of research and development strategies for the improvement of educational practices is a pioneering venture begun in 1964 under the provisions of the Cooperative Research Program of the United States Office of Education and now continuing through Title IV of the Elementary and Secondary Education Act of 1965. Working from a proposed model outlined in Chapter 16, the editors have organized the book around the major components of an instructional system.

Basic to the improvement of education is knowledge about the processes and conditions necessary for efficient learning. In Part I, four authors discuss retention and recall, thought processes, creative thinking, and motivation. A major component of an instructional system is subject matter content and sequence. In Part II, five authors provide scholarly insights into the structures of such disciplines as mathematics, science, English, and reading. They state or imply that the structure of the discipline itself, as formulated by one or more scholars, has provided much of the basis for recent curriculum devlopment.

Instructional materials and media can be considered as the interface between the learner and the subject matter being learned. In Part III, the discussion focuses on instructional television, on some aspects of the status, defects and needs of instructional research, and on computer based instruction. Teaching methods, or the interactions which occur between a teacher and students, have been the subject of many research studies in the last decades. In Part IV, two recent paradigms for research on teaching are presented. One approach discusses what is commonly called micro-teaching and the other is interaction analysis.

Finally, in Part V, three research administrators discuss the current application of research and development strategies for the improvement of educational practices. The present scene in educational R & D is one of rapid change brought about largely by the advent of federal funding. The development of these changes with reference to federal involvement is traced in one chapter. In another chapter the focus and programs of the Wisconsin Research and Development Center for Cognitive Learning are described. And in the final chapter, an author presents an output oriented model of R & D.

Dembar Educational Research Services, Inc. Box 1148
Madison Wisconsin 53701

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

INCREASED COMMUNICATION IN LANGUAGE LEARNING

DIETER M. MESSNER

St. Olaf College
Northfield, Minnesota

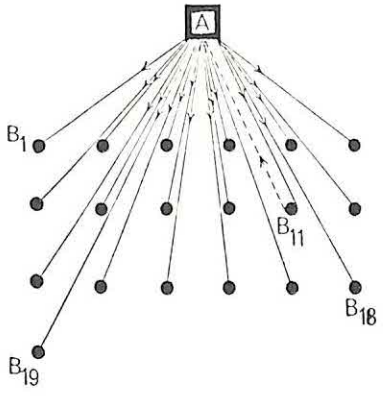
ABSTRACT

There is little opportunity for the individual student to speak in class. The ratio of speaking time to the time of other language activities is improper. The speaking time of the individual can be prolonged many times over in a 2-way communication model, based on the principle of paired learning with planned partner rotation and the teacher acting as a supervisor with special functions. This communication model provides the individual learner with a maximum of speaking time under controlled conditions.

FACILITY IN extemporaneous speaking of a foreign language seems to be an unattainable goal for classroom language learners. The individual student has little chance of conversing in the foreign language. Only two people--the teacher and one student--speak at one time in a real dialogue situation, while the others have to be silent and listen. Because of the usual student: teacher ratio, the opportunity for each student to participate in a dialogue is limited.

DIAGRAM 1

Diagram 1 shows the traditional predominantly one-way communication structure in class, at the moment of a teacher-student interaction ($A \rightleftarrows B_{11}$)



We seem to learn language most economically in conversational communication. Dialogue is considered to be the essence of language, realizing its main functions.

Since spoken language is a motor activity as well as sensory process, the length of the speaking act, the number and frequency of motor stimuli sent from the speech organs to the brain, must be decisive for its learning. The intricate mechanics of speaking cannot be well-learned by merely listening. A certain amount of nerve-cell training and conditioning time is necessary for the learning of any skill.

Let us compare spoken language, a verbal communication skill, with a dual sport. We imagine a beginning tennis player, who, every second day, is given a chance to volley five times with the tennis instructor. When alone he is supposed to learn the game out of a book and by practicing against a cement wall. Let us transfer this situation to the classroom. The instructor stands on one side of the net; the other side is crowded with the multitude of students. Although they have the possibility to learn from an expert, the single learner rarely has the chance to play the "champion" teacher. The learners are not given the chance to play each other. Fortunately, tennis is not learned in such a way as, during practice hours, the "students" do practice with each other.

For the good performance of a skill, the respective chains of nerve-cells in the motor cortex have to be 'tuned in' frequently for a certain length of time. Considering the importance of the length of the speaking process in language learning, the ratio of each learner's speaking time to the time spent on other activities in class, like listening to, reading, or writing the language, is improper. Several language teachers report that the individual student has the chance to participate in classroom conversation for an average of 1 minute or, at best, 2 minutes per lesson. In the ninety lessons of a school year, this would be 180 minutes or only 3 hours of living speech.

The tape laboratory has endowed the foreign language learner with new possibilities; its effectiveness is recognized. For listening and drill purposes it works effeciently, but it offers only "canned language," presents no visibles speaking partner, tires the listener easily, and supplies only audible stimuli.

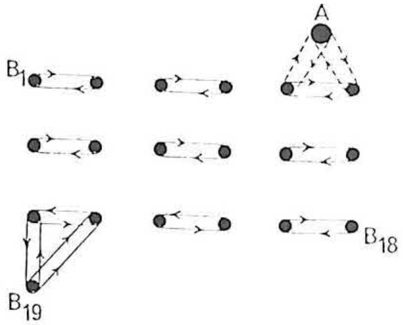
Also, the choral response technique prolongs the speaking time of students and involves the whole class in the activity. But, like all drills, it suffers from a low degree of reality, is sometimes monotonous, and often does not give the student something meaningful to talk about. It would be desirable, to use exercises which involve the learner more personally, while at the same time permit the practicing of the linguistic structures which the teacher wants to drill.

In order to increase the speaking time for the individual learner and to involve him more meaningfully in the foreign language as a life experience, we need a different communication model and learning materials, which can satisfy these needs. I propose that directed exercises in dyads, or pairs of interacting learning partners, equipped with textbook dialogue and lists of questions and answers, made up by the teacher for a certain class, could do the job. Instead of the one teacher facing the many students, students are asked to form pairs, with one possible triad, if the number is uneven. If the actual lesson offers a dialogue, partners A and B begin to read the dialogue alternately. After a possible repetition they continue trying to reproduce as many parts of the dialogue as they can remember. Periodically, the teacher can hand out lists with questions and model answers, and partners A and B can ask each other these questions alternately. These questions and answers are the carriers of the linguistic structures and subject matter which the teacher wishes to review and to reinforce. Questions should refer to personal experiences of the learners or to classroom learning experiences, since meaningfulness seems to be highly desirable for better learning. Partner A is not to read his question off the list, but is to speak it, after he has glanced through the text. For immediate control, the questionnaire furnishes one or more model answers or essential parts of answers (see appendix). If possible, partner B answers the question without looking at the printed material: the answer is approved or corrected and correctly repeated by partner A. This process follows the pattern of stimulus--response--reinforcement. I consider the experience of successful communication to be the reward.

The teacher matches the learning partners and goes among the pairs, acting as an assisting and controlling supervisor and as a resource for correct pronounciation and response. He will suggest new combinations of partners. The rotation of partners creates new motivation and supplies different learning experiences.

DIAGRAM 2

The learning pair or dyads in the classroom. B_{13}-B_{14}-B_{19}-form a triad. A, the teacher, is working with the pair B_5-B_6 at this particular time. In contrast to diagram 1, everyone is interacting in this communication structure.



After a period of directed conversation based on textbook dialogues and question-answer lists, the teacher could suggest continuing with extemporaneous conversation. I have found that the din in the room hardly bothers the students.

Multiple preparatory and immediate controls of the questions and answers possible:

1) preliminary oral preparation in class, 2) preparation by the students at home, 3) immediate control by the learning partner, 4) the model answers on the list, and 5) the assisting teacher.

This communication model could enable us to convert our classroom, whenever desired, into "learning laboratories," where every student is active at the same time and uses his time most profitably. Partner learning, with the ideal ratio of one-to-one, provides maximum participation for each (theoretically 50 percent of the given time); it gives everyone a chance to speak and no one a chance to hide. While the traditional communication structure in a class of twenty students would permit, for example, fifty interactions within a given time, the dyadic model could provide 500 interactions within the same time. My experiences with this technique suggest that it may be more effective to devote an entire class period to it now and then instead of devoting short intervals to it every period. I dedicate every third class period entirely to this method. The "speaking program" for this hour is written on the blackboard: which dialogues the students are to practice, which pages of the question-answer sheets they are to go over and how often, and last but not least, a topic for extemporaneous conversation is suggested.

Since I cannot point to exact experimental results, the hypothesis has the status of an assumption. Classroom observations do, however, confirm its effectiveness. The learners soon gain self-confidence in speaking. The "pathological" symptoms of people learning to speak a foreign language, hesitation and halting speech, disappear soon. Partners can discuss common difficulties and can help each other. It is plausible that especially our late-starting language learners in high school and college should be given this opportunity for facilitating motor activity. Partner rotation provides the advantages of the larger group: a wider scope of individual responses and different learning experiences. The classroom is now more social as compared to the traditional centralized and authoritarian communication structure, which is mainly one-way. There is a more personal contact with the teacher, who meets and works with only two students at one time. The number of students has little importance in this communication model, since any number can be broken down into pairs and one possible triad. This technique is analogous to the communicative verbal learning process in everyday life. The question-answer process is not a completely real dialogue, yet a great deal of any dialogue is devoted to the asking and answering of questions. It has been said that the ability to ask and answer questions efficiently lies at the root of communication. The sets of questions and corresponding model answers can be considered as the smallest units of split-up dialogues. There is an added responsibility for the teacher who does have to make up his own questions. On the other hand, good class rapport can be promised, based on my own experiences.

Besides the compulsory written final exam, an oral final is scheduled for each student, in order to boost the motivation for the learning of the spoken language. Enough questions and answers accumulate during one semester to make a sufficient basis for a meaningful test-dialogue.

The question-answer process, as the interaction between two people, is the basis of programmed instruction. It is my hope that in the near future we will have a "teacher-proof" textbook for dyadic language learning, which will incorporate meaningful conversational question-answer interaction, important linguistic structures, as well as the most frequently-used vocabulary of the target language. Such a book would be practical in schools and could be used for the self-instruction of learning partners or for single learners, if tapes were provided.

APPENDIX

A: What is a traveler asked at the border ?
B: He is asked: "Do you have anything to declare?"/"May I see your passport?"/"Have you a vaccination certificate for the dog?"
A: (repeats or corrects)

A: Have you already been abroad?
B: Yes, already several times. /No, not ever.
A: (repeats or corrects)

A: Was there any mail for you yesterday?
B: Yes, I got two letters and one postcard.
A: (repeats or corrects)

A: Do I stand opposite you?
B: No, you are sitting next to me.
A: (repeats or corrects)

A: Where would you travel, if you had time and money?
B: If I had time and money, I would go to Florida or Switzerland.
A: (repeats or corrects)

A: Do you always smoke a cigarette after dinner?
B: No, I don't smoke at all. / Just sometimes.
A: (repeats or corrects)

A: What do you do when you feel dirty?
B: When I..., I take a bath.
A: (repeats or corrects)

A: Through what did Dr. Christian Barnard become famous?
B: Dr. Barnard became famous through his heart-transplant.
A: (repeats or corrects)

A: What will you do in the coming summer?
B: I will work in order to earn money. / I will attend a summer school.
A: (repeats or corrects)

A: In what month were you born?
B: I was born in August.
A: (repeats or corrects)

A: What can one eat with a spoon?
B: With a spoon one can eat, for example, soup, ice cream, and pudding.
A: (repeats or corrects)

A: With which words does one begin a letter?
B: One begins a letter with the words: Dear Sir, or Dear Madame.
A: (repeats or corrects)

A: In the summer, did you live with your parents or with your grandparents?
B: In the summer I lived with an aunt. / I lived in a camp.
A: (repeats or corrects)

A: Where do you get your hair cut?
B: I get my hair cut at the barber shop. /... at the hairdresser.
A: (repeats or corrects)

A: Do you like to dance?
B: No, I am not a dancer. / I never learned it. / Yes, especially the English waltz and the rumba.
A: (repeats or corrects)

A: In the winter, what does one put into the cooling water?
B: One puts antifreeze into the cooling water.
A: (repeats or corrects)

A: What disadvantages does a trailer have?
B: It is difficult to maneuver on mountain roads, it is often illegal to park and it is subject to many restrictions.
A: (repeats or corrects)

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

TEACHER PLANNING OF INSTRUCTIONAL MATERIALS FOR LOWER-CLASS CHILDREN

GEORGE MOVESIAN
Herbert H. Lehman College

ABSTRACT

The purpose of this research project was to investigate the possibility that, given a free choice of materials, elementary school teachers would not differentiate between materials judged appropriate for lower-class children and materials judged appropriate for middle-class children. The subjects were twenty-eight and twenty-nine teachers of lower-and middle-class children, respectively. Teachers were selected from a random list of schools identified as either lower- or middle-class by the respective superintendents, using Alba Edwards, Alphabetical Index of Occupations and Industries as a guide. When teachers of lower- and middle-class children were compared in selecting material for children in the respective classes, no significant differences were found. The failure to reject the two basic subhypotheses seems to support the possibility that, given a free choice of materials, elementary school teachers of lower-class children do not differentiate between materials judged appropriate to be used by middle-class children.

IT HAS become an educational axiom that the good teacher is one who is aware of individual differences among the students. Some of these differences result from differing social class backgrounds. Each social class has a unique sub-culture which affects student behavior in school--in his modes of perception, reaction to frustration, etc. Hence, effective teaching demands provision for individualizing instruction, based on an awareness of the characteristics of various class cultures. If one has a group of children with predominantly lower-class backgrounds, he selects suitable materials, i. e. familiar, recognizable, and within the cultural context of the student. A progression from the known to the unknown is essential. Most teachers know this. Almost all have been exposed to such ideas in courses in education. However, do teachers use this knowledge? Do they plan lessons which indicate an awareness of the social background of the students? One clue to the answers to these questions may be discovered in the response to the question: "When given a free choice, do teachers select materials appropriate for students in terms of their backgrounds?"

It seems safe to assume that if teachers selected the appropriate materials there would be some opportunity for them to reflect and adapt their teaching methodology to specific backgrounds of their students.

This study attempts to investigate this question. Accordingly, the following hypotheses were tested:

When given a free choice of materials, teachers in schools of predominantly lower-class children will select materials that would not be significantly different from materials selected by: (a) teachers in middle-class schools and (b) independent judges.

Second, when given a free choice of materials, teachers in schools of disadvantaged children will select materials judged to be more appropriate for the background and experiences of middle-class children then those of lower-class children.

To test these hypotheses, the following assumptions were made: First, different social classes exist within our American culture and school children reflect, in various ways, the family class position. Second, the middle class has a greater influence upon the curriculum of the schools than any combination of the other divisions of the social classes. Third, the principle of teaching children by utilizing their experiential background to help make the teaching-learning activity meaningful to the learner is educationally sound. Fourth, the principle of meeting individual needs, wants, and interests in the planning of teaching-learning activities is educationally sound.

SAMPLE

Sixty-one teachers were selected from a random list of schools identified as either lower- or middle-class by the respective superintendents, using Edwards, Alphabetical Index of Occupations and Industries (1) as a guide. The social class position of the student population of each classroom was carefully investigated. Only those teachers whose student populations contained at least 60 percent lower- or middle-class were used. Lower-class children were

TABLE 1

A COMPARISON OF MIDDLE CLASS MATERIALS SELECTED BY TEACHERS OF LOWER- AND MIDDLE-CLASS CHILDREN

Middle-class Materials		40-55 %*	56-66 %	67-100 %	Totals
Teachers of Lower-class Children	obsvd.	5	12	11	28
	exped.	(6.88)	(11.79)	(9.33)	
Teachers of Middle-class Children	obsvd.	9	12	8	29
	exped.	(7.12)	(12.21)	(9.67)	
		14	24	19	57

$x^2_{2\,df} = 1.6047$ ∴ Fail to Reject

*43 % was the lowest percentage of middle-class materials selected by both groups of teachers.

those children whose parents, when gainfully employed, were classified as unskilled. Middle-class children were defined as those children whose parents, when gainfully employed, were classified as professional, proprietary, or managerial, except farm. The investigator obtained information from twenty-eight teachers of lower-class children and twenty-nine teachers of middle-class children. [1]

The teachers were located in five urban and three suburban communities. The eight separate school districts were within 25 miles of the center of a large city.

INSTRUMENT

In attempting to have a common subject area from which teachers would plan a unit of work, a unit on transportation was selected because it is required by the New York State Department of Education. To allow for maximum freedom for teachers to select instructional materials for a unit of work, a resource unit titled "Transportation" for elementary-school children was constructed. This resource unit consisted of a carefully planned series of suggestions centered on the topic and organized with ideas, materials, and procedures to help the teacher in preplanning the learning unit. The recommended books were measured by a jury of three judges according to the Movesian Reading Rating Scale of Class Culture which rates books for their appropriateness for children of Lower- and middle-class backgrounds. [2] The selection of the jurors was based upon their respective background which was consistent with this study, i.e., that they understand the problem of identifying instructional materials for lower-class children and are interested in the improvement of instruction for these children.

At a personal meeting with the teachers, the resource units were distributed. The teachers were asked to plan a unit of work for their respective classes. At no time was social class mentioned.

Each teacher had the freedom to select as many items from the resource unit as he thought sufficient for a teaching unit. Therefore, the number of books and specific titles selected varied from teacher to teacher.

RESULTS

For purposes of comparing the selections obtained from the teachers of lower- and middle-class children, a percentage of the total materials (books only) defined as middle-class was computed. The percentage score for each teacher was placed in the appropriate cell of a two-by-three table (see Table 1). The Chi-square test of association was used to test the hypothesis of "no difference" in the materials selected by teachers of lower- and middle-class children, respectively. The hypothesis, which stated that given a free choice of materials, teachers in schools of predominantly lower-class children selected materials that are not significantly different from materials selected by teachers in schools of predominantly middle-class children, was not rejected. The 95 percent confidence level was used for all tests of significance.

Each judge independently rated the books according to Movesian Reading Rating Scale of Class Culture. Only those materials agreed upon by the entire panel of jurors to be on the same side of the scale were included in the unit. Thus, all three judges agreed that, in their opinion, the item was either lower- or middle-class oriented prior to its inclusion in the final instrument.

To test the subhypothesis, the data was treated in order to compare the differences of means: i.e., a comparison of the scale means of the judges and the teachers of lower-class children. The scale means for each of the entire listing of ninety-six books were computed from the average score assigned to each book by the judges. The mean of the scale scores of all ninety-six books included in the unit was computed to be 4.962. This was then considered to be the mean of the scale scores for the judges. The t-test of Independent Samples was used (see Table 2).

TABLE 2

A COMPARISON OF MEAN SCALE SCORES OF THE JUDGES AND TEACHERS OF LOWER-CLASS CHILDREN

	Mean of the Scale Scores	Computed t-value	Degrees of Freedom	F-Test Variance	N_1	N_2
Teachers	4.720	0.9153	29	6.974	28	3
Judges	4.962					

t_{29df} H_A: $p_1 - p_2 = 0$ $\therefore$ Fail to Reject, $\sigma = .26457$

F-Test H_A: $p_1 - p_2 = 0$ $\therefore$ Fail to Reject

To determine the validity of the assumption that the variance of scores in the populations are equal to the t-test of independent samples, and F-test of variance was required. The probability with which the distribution could have occurred from populations with identical distributions was greater than 0.05 for both the t-test of independent samples and the F-test of variance; therefore the hypothesis of "no difference" between the selection of materials by teachers of lower-class children and the judges was not rejected. Because the resource unit had 66 percent of the total listing of books identified as middle-class materials, and the t-test of "no difference" between the teachers of lower-class children and the judges was not rejected, the findings supported the second hypothesis. That is, given a free choice of materials, teachers in schools of predominantly lower-class children select materials judged more appropriate to the background and experiences of middle-class children than those of lower-class children.

The teachers of lower-class children chose a greater percentage of materials judged appropriate for middle-class children than the teachers of middle-class children (see Table 1). The mean scale score of all the books selected for each group was compared (see Table 3). The hypothesis of "no difference" was not rejected for either the t- or F-tests.

Although the teachers of lower-class children selected a greater percentage of the materials judged to be more appropriate for middle-class children than teachers of middle-class children, the Movesian Reading Rating Scale of Class Culture score of these selections was below those of the teachers of middle-class children. The teachers of middle-class children selected more books with a higher scale score than the other group of teachers.

DISCUSSION

Teachers not only seem to ignore or are ignorant of lower-class materials, but also do not seem to acknowledge or accept lower-class culture. The refusal of teachers and schools to accept much of the language and overt behavior of lower-class children, concomitant with the selection of materials more appropriate for the middle-class child's background, suggests that teachers do not accept nor acknowledge lower-class culture.

Reluctance on the part of teachers to use appropriate materials for lower-class children may also stem from the following reasons: (1) It may be that the teachers feel lower-class children should aspire to middle-class values. They may judge that this "Rising" by using the materials which were deemed appropriate for middle-class children. (2) It may

TABLE 3

A COMPARISON OF MEAN SCALE SCORES BY TEACHERS OF LOWER- AND MIDDLE-CLASS CHILDREN

	Teachers of Lower-Class	Teachers of Middle Class	Computed t-value	F-test of Variance	N_1	N_2
Mean Scale Score	4.7195	4.7662	.345 $\sigma = .135$	1.383	29	28
55df	N 28	N 29	H_A: $p_1 - p_2 = 0$ $\therefore$ Fail to Reject	H_A: $p_1 - p_2 = 0$ $\therefore$ Fail to Reject		

be that teachers feel social class membership plays an insignificant role in growth and development, especially in the area of instructional materials, i. e., books for children. In essence, teachers do not appear to be meeting the lower-class child's need for instruction geared to individual differences based on social class.

CONCLUSIONS

From the data tabulated, it was found that when given a free choice of materials, teachers in schools of predominantly lower-class children selected materials that are not significantly different from the materials of (a) teachers in schools of predominantly middle-class and (b) the jurors rated and preselected.

Second, when given a free choice of materials, teachers in schools of predominantly lower-class children select materials judged more appropriate for the background and experiences of middle-class children than those of lower-class children.

These conclusions suggest that further research be undertaken testing the following hypotheses:

(1) When given a free choice of materials, elementary school teachers of lower-class children do not differentiate between materials judged appropriate to be used by lower-class children and materials judged to be used by middle-class children, or deliberately ignore using such materials.

(2) It seems that teachers of lower-class children did not appear to meet the lower-class child's needs of individual differences on a social class basis in the pre-planning of instructional materials.

(3) Materials which could be judged appropriate for lower-class children were few when compared to materials which could be judged appropriate for middle-class children.

FOOTNOTES

1. Four returns were deemed unusable. Two teachers of lower-class children refused to complete the instrument after agreeing to participate. One teacher of lower-class children became seriously ill and took a leave from his teaching position. Because one teacher of middle-class children selected all ninety-six books, the teacher really did not discriminate in his selection of books. Thus, the information gathered came from fifty-seven teachers.

2. Because the criterion of occupations of parents was used to identify the social class position of the children for whom the unit was intended, the same criterion, i. e., occupations, was used as the major basis to determine which instructional materials were appropriate for lower- and middle-class children. Because most materials investigated had some overlapping of social class orientations (in terms of occupations), the jurors were asked to evaluate each of the materials in terms of a 7-point scale. The numbers one, two, and three were used for those materials judged to be appropriate for lower-class children. The numbers five, six, and seven were used for those materials judged appropriate for middle-class children. Any materials which were judged as equally suitable for both lower- and middle-class children (number four) were not included in the resource unit. Only those materials agreed upon by the entire panel to be on the same side of the scale were included in the unit.

REFERENCE

1. Alba Edwards, Alphabetical Index of Occupations and Industries, U. S. Department of Commerce, Washington, D. C., 1950.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

SPECIALIZED STUDY SKILLS, DEVELOPMENTAL-READING INSTRUCTION AND COUNSELING

R. N. SAWYER[1] and L. W. MARTIN[2]

University of Missouri-Rolla

ABSTRACT

Ninety beginning freshmen at the University of Missouri-Rolla were utilized as subjects. The sample was sub-divided into three equal groups; the experimental group was composed of pre-registered students enrolled in an intensified instructional program in Study Skills and Developmental Reading and were provided specialized counseling services; one control group was composed of students that had taken part in the pre-registration program at UMR, but were not enrolled in the Study Skills and Developmental Reading instruction and were not provided specialized counseling services; and the other control group was accorded no variation in traditional method of enrollment or instruction. All three groups were matched on their obtained School and College Ability Test scores, English and Mathematics Placement Test scores, college majors and semester hours attempted. All students in these groups were above average in quantitative skills and below average in verbal skills.

The experimental group met for developmental reading and study skills instruction on a bi-weekly basis for no academic credit. The study skills instruction included material dealing specifically with both study habits and study attitudes. The developmental reading sequence consisted of both machine and workbook instruction. The counseling service was provided on a voluntary basis and was initiated at the time of pre-registration. A statistically significant difference in grade point average (GPA) was discerned after the first semester of the project. While no significant differences in GPA existed after the second and third semesters of the investigation, the attrition rate was considerably lower for the experimental group than for the control groups. Highly significant gains were evidenced in reading rates with no significant loss in reading comprehension.

PRESENTLY, the demand for competent scientists and engineers is increasing while college enrollment in these scientific and technological areas is decreasing. One of the resulting challenges placed upon institutions serving the technological clientele is to provide additional special student services that will help insure graduation of the largest possible number of well-educated and socially adjusted scientists and engineers.

In a report (1) entitled, "Factors Influencing Engineering Enrollment," published by the American Society for Engineering Education and supported by the National Science Foundation, the statement was made that"... engineering enrollments over the country are down... despite the coincidental rise in the college population." Further, this report recommends that

> ... a major effort be made to insure that a larger fraction of the students who enter engineering successfully complete the degree requirements. If engineering attrition were reduced to the level of 1950, the effects would be the same as though freshmen enrollment were increased by almost 29 percent. In 1950, one hundred were required to graduate about 63 engineers; in 1959, approximately 129 freshmen were required to produce 63 engineering graduates.

After considering the attrition among first semester engineering freshmen at the University of Wisconsin (2) in 1963, a conclusion was that "programs for remediation of dificiencies in previous training such as remedial study skills clinics might be undertaken by engineering colleges desiring to minimize attrition of prospective engineers." This statement further supported the need for a developmental reading and study skills program at the University of Missouri-Rolla (UMR).

Numerous psychologically oriented counseling centers have been provided to assist students with their educational, vocational, and personal planning and adjustment. Fewer colleges and universities have provided specialized instruction in study skills and developmental reading for their students. Only a limited number of institutions have offered a combined program of study skills and developmental reading instruction with psychological counseling for their students. Because few institutions of higher education have as homogeneous a student body as that at the University of Missouri-Rolla, the researchers desired to investigate the effects of a combined program of study skills and developmental reading instruction with psychological counseling upon a selected sample of engineering and science students.

PROBLEM

The purpose of the investigation was threefold:

1. to ascertain the effect of a specialized program of study skills, developmental reading, and psychological counseling on the obtained GPA of the experimental group;
2. to ascertain the degree of success associated with a developmental reading program when employed with an experimental group manifesting limited verbal skills; and
3. to ascertain the effect of a specialized program of study skills, developmental reading and psychological counseling on the attrition rate of the experimental group.

SAMPLE

Ninety beginning freshmen students at the University of Missouri-Rolla were utilized as subjects. The sample was sub-divided into three groups of thirty students. The Experimental Group represented freshmen students that had participated in the UMR Summer Pre-Registration Program, 1966, and had volunteered for enrollment in the one hour, no credit, bi-weekly study skills and developmental reading program. Psychological counseling was available to the students as was felt necessary or desirable by the students and their instructor. The Control Group I represented freshmen students that also had participated in the Summer Pre-Registration Program, 1966, but were not enrolled in the study skills and developmental reading program nor provided counseling. The Control Group II represented freshmen students that had not participated in the Summer Pre-Registration Program, 1966, were not provided specialized instruction, nor became involved in counseling.

The three groups were matched in terms of their declared major field of study and semester hour course load. The sample included ceramic, chemical, civil, electrical, mechanical, and metallurgical engineers as well as chemistry, geology, and mathematics students. The ninety subjects were also matched on the basis of their obtained quantitative and verbal scores on the Cooperative School and College Ability Test (SCAT). Further, the subjects were matched on the basis of scores earned on the Missouri College English Test(MCET) and Missouri Mathematics Placement Test (MMPT).

The analysis of variance (3) procedure was employed at the beginning of the project to discern if any statistically significant difference existed among the three groups on the obtained SCAT, MCET, and MMPT scores as well as semester hours attempted. No significant difference was found at the . 05 level.

METHODOLOGY

Ninety subjects were selected for participation in the investigation at the time of pre-registration or at the time of fall registration. The enrollment of all participants in the study was verified after the end of the late registration period. In two cases, the students selected for participation in the Experimental Group and Control Group I did not arrive on campus to begin their class work. The criteria for the selection of the two substitute students was identical to the criteria employed in the selection of the original sample. Both control groups were selected in final form just after the close of the registration period.

Subsequent to the termination of the pre-registration program, students selected to compose the Experimental Group were contacted by letter and informed of the objectives of the project. Further, it was the intent of the investigators to provide a bridge for the students to assist in their transition from the high school academic setting to the university academic setting.

The class meetings were scheduled for two hours weekly (T and Th) and the participants were to receive no academic credit. A course outline was provided and followed closely during the entire project. Counseling appointments were scheduled as the participants or the instructor felt counseling was needed.

While the minute details of the methodology will not be discussed, the present investigators desire to point out that all reading test data were collected during the initial phases of the project and during the termination phase of the project. These data were obtained by employing the pre-post test procedure with alternate forms of a diagnostic reading examination. The basic reading materials employed in the project were Miller's Increasing Reading Efficiency (4) and Maintaining Reading Efficiency (5) workbooks. This basic approach was supplemented with various mechanical devices. In addition, Robinson's Effective Study (7) textbook provided the basic frame of reference for the intensified instruction in study skills and study habits.

The first semester GPA and semester hours passed were obtained subsequent to the end of the first semester. This period of time was allotted to insure that any grade changes had taken place. During the following two semesters, data relative to the participants' GPA's and attrition rates were collected and analyzed.

The analysis of variance technique (3) was utilized to investigate the differences between mean semester GPA's of the three groups. Further, the same statistic was employed to ascertain the differences in mean semester hours attempted and completed by the experimental and both control groups. The critical region was established at the . 05 level.

The "student's" t statistic (6) was employed to discern the differences of the pretest mean reading scores (reading rate, comprehension, and efficiency) and posttest mean reading scores. The critical region was established at the . 01 level.

RESULTS

The primary purpose in this study was to investigate the relationship between specialized instruction in study skills, developmental reading and counseling and academic progress. Secondary objectives in the study were the following: (1) to investigate the degree of success associated with the developmental reading procedures when utilized with a highly scientific and techologically oriented sample; and (2) to initiate an investigation of the attrition rate of the experimental and control groups.

Contained in Table 1 are the analysis of variance summary tables for difference between mean semester hours attempted and semester hours completed for the Fall, 1966, Spring, 1967, and Fall, 1967

TABLE 1

ANALYSIS OF VARIANCE FOR THE DIFFERENCES AMONG SEMESTER HOURS ATTEMPTED AND SEMESTER HOURS COMPLETED

Semester	Variable	Source	df	MS	F
Fall, 1966					
	Semester Hours Attempted				
		Treatment	2	3.82	1.82
		Error	87	2.10	
	Semester Hours Completed				
		Treatment	2	21.24	2.43
		Error	87	8.73	
Spring, 1967					
	Semester Hours Attempted				
		Treatment	2	12.59	1.01
		Error	83	12.45	
	Semester Hours Completed				
		Treatment	2	39.97	2.16
		Error	83	18.54	
Fall, 1967					
	Semester Hours Attempted				
		Treatment	2	.01	.002
		Error	66	3.36	
	Semester Hours Completed				
		Treatment	2	.59	
		Error	66	20.47	

Not significantly different from zero at the .05 level

semesters for the experimental and control groups.

Presented in Table 2 are the analysis of variance summary tables for the difference between mean semester GPA's for the Fall, 1966, Spring, 1967, and Fall, 1967, semesters for the experimental and control groups.

TABLE 2

ANALYSIS OF VARIANCE FOR THE DIFFERENCES AMONG OBTAINED GRADE POINT AVERAGES

Semester	Source	df	MS	F
Fall, 1966				
	Treatment	2	1.48	3.53*
	Error	87	.42	
Spring, 1967				
	Treatment	2	.47	.77
	Error	83	.61	
Fall, 1967				
	Treatment	2	.87	1.31
	Error	66	.67	

*Significantly different from zero at the .05 level

Developmental Reading

Table 3 contains the means and standard deviations of the experimental group for the pretest and posttest reading rates, reading comprehension, and reading efficiency scores. Also presented are the resultant "student's" t values from the statistical analysis regarding the test of equality between pretest and posttest mean reading scores.

Attrition Rate

Indices of the attrition rate among the three groups are contained in Table 4. The numbers are based on the enrollment figures at the start of each successive semester.

DISCUSSION

The analysis of variance (Table 1) of differences between mean semester hours attempted and semester hours completed by the experimental and control groups suggests that there was no significant disparity between the groups for either variable. However, this analysis provides no data regarding the level

TABLE 3

MEANS, STANDARD DEVIATIONS, AND "STUDENT'S" t VALUES FOR TEST OF EQUALITY BETWEEN PRE- AND POST- READING SCORES

Variable	Pre	Post	"t"
Rate in Words per Minute			
Mean	196. 79	734. 71	19. 66*
SD	40. 81	456. 44	
Comprehension in Percentage			
Mean	80. 07	73. 96	2. 07
SD	11. 46	10. 61	
Efficiency in Words per Minute			
Mean	156. 54	514. 69	6. 60*
SD	33. 51	284. 95	

* Significantly different from zero at the . 01 level

TABLE 4

NUMBERS OF STUDENTS THAT HAVE ENROLLED DURING THE THREE SUCCESSIVE SEMESTERS

	Experimental	Control I	Control II
Fall, 1966	30	30	30
Spring, 1967	28	29	29
Fall, 1967	26	23	20

at which the three groups completed the course work attempted.

When the semester GPA's obtained by the three groups were submitted to an analysis (Table 2), a significant difference was discerned for the initial semester of the study and no significant differences for the final two semesters. Further analysis of the first semester GPA's by the Tukey (3) method for multiple comparisons indicated a difference between the Experimental and Control II groups and apparently accounted for the rejection of the hypothesis of mean equality. The fact that a significant difference in GPA's was discerned during the initial semester would suggest that the time spent in special instruction and counseling was valuable during the initial semester of enrollment. Further, it may suggest that the experimental group progressed more rapidly with their adjustment and transition from high school. Also discerned from the data was that the experimental group maintained a constant level of achievement and the two control groups improved to a level similar to the experimental group. This was especially the case during the third semester of enrollment when many of the students in the control groups had terminated their education. In essence, the withdrawal of the lower achieving students tended to make the groups more homogeneous in terms of GPA.

Vitally important to all engineering educators was the attrition rate of the three groups. These data were presented in Table 4. The fact was noted that the Control Group II had been reduced by 33 percent while the experimental group suffered only a 13 percent loss. Assuming these attrition rates will tend to level off as the participants approach graduation, this experimental approach would help considerably to assist engineering students in successfully completing their education once initiated.

The data presented in Table 3 strongly suggested that the reading phase of the experimental project was highly successful, both in terms of increased reading rate and no significant decrease in the level of comprehension. Further, the efficiency scores indicated a high positive change which implies a more effective and efficient reading process on the part of the experimental group as a result of the reading instruction. Discerned was the high level of comprehension at the onset of the project and the below average reading rate. The rate was increased nearly 300 percent without a significant loss in comprehension level. It would appear that the below average verbal skills of the experimental group did not hinder their reading growth. The students seemed to be highly motivated to become involved in the reading program. They saw the reading skills as vitally important, both in terms of the present academic program and as an avenue for keeping abreast with a rapidly changing technological field.

It seems reasonable to conclude that the experimental project was successful, both in terms of GPA and reduced attrition rate. Further, it would appear that a developmental reading program can assist students with limited verbal abilities to progress toward more effective and efficient reading skills.

FOOTNOTES

1. R. N. Sawyer is Associate Professor of Psychology at the University of Missouri-Rolla.
2. L. W. Martin is Director of Institutional Studies at the University of Missouri-Rolla.

REFERENCES

1. Engineering College Administrative Council, American Society for Engineering Education, Factors Influencing Engineering Enrollment, October, 1965, pp. 1- 5.
2. Greenfield, L. B., "Attrition Among First Semester Engineering Freshmen," Personnel and

Guidance Journal, 42:1003-1010, June, 1964.

3. Guenther, William C., Analysis of Variance, Prentice-Hall, Inc., Englewood Cliffs, 1964, pp. 3-63, 174.

4. Miller, Lyle L., Increasing Reading Efficiency, Holt, Rinehart, and Winston, New York, 1964.

5. Miller, Lyle L., Maintaining Reading Efficiency, Developmental Reading Distributors, Laramie, 1966.

6. Ostle, Bernard, Statistics in Research, Iowa State University Press, Ames, 1963, pp. 119, 528.

7. Robinson, Francis P., Effective Study, Harper and Brothers, New York, 1962.

You Are Invited to Help Launch a New Professional Quarterly

ENVIRONMENTAL EDUCATION is communications aimed at producing a citizenry that is knowledgeable concerning our biophysical environment and its associated problems, aware of how to help solve these problems, and motivated to work toward their solution. How better to achieve this pervasive ecological conscience is the focus of this professional Journal.

What This Journal Is

Environmental Education is a new quarterly devoted to discovery and dissemination in the emerging field of multidisciplinary conservation communications. Through a recording of insight and information about pioneer research and interpretation projects and programs, this Journal will act as a catalyst in encouraging and exchanging effective studies and developments toward broad public ecological awareness, understanding, and action.

Whom It's For

This is a professional Journal for all those who are beginning systematically to study and practice the new interpretation of America's resources—their essential integrity, their inexorable relationships with modern man, their beauty, their bio-physics, their utility, their socioeconomics, their conservation: for researchers in education and communications, for conservation educators, for environmental designers, for interpretive naturalists, for regional planners, for resource investigators and managers, for practitioners of public relations in resource industries, for conservation information specialists, for outdoor writers, for lay leaders in land and water programs—for all those concerned with discovering and applying better ways of developing a viable "conservation conscience" on the part of America's citizenry.

Where

Environmental Education is published four times a year by Dembar Educational Research Services, Inc., Madison, Wisconsin, the publishers of *The Journal of Educational Research* and *The Journal of Experimental Education.* It will initially be 32 pages, 7x10; priced at $2 a copy, $7.50 a year, $12 for two years; student rates will be $4.75 a year.

Inscribe Your Name Today

DEMBAR EDUCATIONAL RESEARCH SERVICES, INC., BOX 1605, MADISON, WISCONSIN 53701

_____ **I want to be a sponsor. Please enter my subscription to *Environmental Education* for two (2) years at $12.**

_____ **I'd like to try *Environmental Education* for a year. Please enter my subscription at $7.50.**

_____ **I'm interested in *Environmental Education*. Please send me a copy of the current issue for $2.**

_____ **I am a student. Please send me *Environmental Education* for a year at $4.75.**

_____ **I enclose a manuscript for your consideration.**

_____ **I enclose a check.** _____ **Bill me. Bill** _____

NAME _____

ADDRESS _____

CITY _____ **STATE** _____ **ZIP** _____

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

CONFIDENCE LEVEL AND SIGNIFICANCE LEVEL: SEMANTIC CONFUSION OR LOGICAL FALLACY

WILLIAM M. STALLINGS and SUSHILA SINGHAL
University of Illinois

ABSTRACT

Despite repeated warnings against its usage, researchers and textbook authors continue to employ the phrase "level of confidence" as a synonym for the phrase "level of significance." Three possible sources of this incorrect usage are noted: (a) imitation; (b) the related but different concepts of fiducial inference and confidence interval estimation; and (c) the myriad of terms associated with confidence. It is recommended that the magnitude of level of confidence be restricted to that of the complement of the level of significance and also that the term level of confidence should be used only in connection with interval estimation.

MOSTELLER, Rourke, and Thomas, in a text largely based upon material used on the popular NBC "Continental Classroom" course in probability and statistics, have made the following distinction between level of confidence and level of significance.

> In a testing problem, we speak of the significance level of the test. It measures the probability of rejecting the null hypothesis when it is true. It is the complement of the confidence level, if a confidence level is used to make the test. Thus, if the confidence level is chosen as 0.95, the significance level is 0.05 (16:304).

Somewhat differently, McNemar has cautioned that "the term 'level of confidence' should not, as it frequently is, be misused in place of 'level of significance.' The first term pertains to interval estimation, the other to hypotheses testing (15:93)."

Consequently, while reading a recent elementary statistics textbook (14), we were jarred by the repeated use of the phrase, "5 [or 1] percent level of confidence"--when the author explicitly was referring to the critical region. Clearly, level of confidence was being used as a synonymous expression for level of significance.

Once sensitized to this questionable equating of expressions, we began to discover the "5 percent level of confidence"--or 10 percent or 1 percent--cropping up in quite varied and sometimes unexpected places. An article in Educational Horizons (13:103), another in the Journal of Educational Measurement (20:111), and a handful of texts, uncovered by a limited search of the literature, provided further examples of this usage. Apparently, the use of the 5 percent level of confidence is unrelated to prestige; it is employed by Edwards (6:242) and by Guilford (9:152). Moreover, date of publication does not seem to be a factor in the use of this terminology (e.g., 11:26 and 12:13). Hence the equating of level of confidence with level of significance appears to be widespread and of long standing.

Although the confusion of level of significance with level of confidence has been noted before (e.g., 3 and 18), there has been little speculation about the origins of that confusion. Three explanations seem plausible. It may be that statisticians, like everyone else, learn by imitating exemplars. Hence, the first explanation, and perhaps the proximate cause, is the equivalent use of the two terms in some textbooks and journals--and even in a dictionary.

For example, consider this definition from a dictionary of psychological terms.

> Confidence level specifies the percentage of samples that may be expected to fall outside the confidence interval: one speaks of a statistic as stable or significant at the 5 percent or 1 percent level. Although this meaning is rigorously definable, it is confusing; 5 percent is not the percent in which we are confident; it is the percent that we must risk (7:206).

Suggestively, much of the English and English (7) discussion of confidence is found under the rubrics of fiducial limits and fiducial probability. And here we arrive at the most likely explanation for the confidence level vs. level of significance confusion--namely, the related, but different concepts of fiducial inference and confidence interval estimation.

Although we do not intend to reopen the once acrimoniously debated--and yet unsettled--question concerning the relationships between fiducial inference and confidence interval estimation, it might prove worthwhile to revisit those concepts. First, what is meant by fiducial inference? In a statistical dictionary the following definition is offered:

> Fiducial Inference: A type of statistical inference based on fiducial distribution introduced by R. A. Fisher. The object of the inference is to make probabilistic statements about the values of unknown parameters and to that extent it resembles the theory of confidence intervals. In simple cases results from fiducial theory agree with those from the theory of confidence intervals, but this is not so in general. When this type of inference and the related confidence-interval type were first propounded, they were often confused. The confusion survives to the extent that "fiducial" is sometimes applied to inference of the confidence-interval type (10:108).

It is doubtful whether this short definition can illuminate the murky concept of fiducial inference. However, these two definitions do reflect the historical confusion of fiducial and confidence interval inferences.

Now that we have touched on the elusive notion of fiducial inference, let us briefly reexamine some statistical competitors: confidence, confidence intervals, and confidence limits. After all, most applied statistics textbooks do pay homage to the Neyman-Pearson concept of confidence. An excellent definition of confidence is provided by Chandler:

> Confidence, a concept customarily reserved for discussions of interval estimation, is the faith which one is willing to place in a statement that an interval established by a sampling process actually contains or bounds a parameter of interest. One generally expresses this faith statistically by affixing to each interval a confidence coefficient, or confidence probability, which can be written as $1 - \epsilon$, where $\epsilon = p/100$ for $0 \leq p \leq 100$ and p is usually taken to be a very small number (3:429).

As is well known, the confidence coefficient is not the usual relative frequency probability. The probability of a given confidence interval containing the parameter under consideration is either zero or one. The confidence coefficient does represent a degree of faith in the particular procedure for constructing confidence intervals, that over a large number of trials with equal size random samples, a given per cent of the intervals so constructed will capture or contain the requisite parameter.

However, the meaning and interpretation of confidence apparently are concepts under which the logical underpinnings are not as strong as is usually implied by applied statistics texts. A recent book review in Science underscored this point:

> The deeper difficulties of the theory [Neyman-Pearson confidence interval estimation] appear when it is recognized that the interpretation of individual sets of data is after all a necessity. A theory which deals only in long-run frequencies does not possess in itself the resources to deal with such interpretation (5:663).

None the less, most of our textbooks in applied statistics reflect the historical legacy of both the Fisherian and Neyman-Pearson schools. In these texts significance tests and null hypothesis lie comfortably along side the two types of errors and confidence. Moreover, even most authors of applied statistics texts admit that the statistical logic of fiducial probability and confidence differs. But with rare exception (e.g., 19:113) these same authors dismiss fiducial limits and fiducial probability (if mentioned at all) with a footnote such as: "Strictly speaking, these are 'fiducial limits' and not confidence limits. For practical purposes the distinction is so slight as to be not worth making, and because of its greater intelligibility the term confidence limit is used here (2:34)." To recapitulate, it has been shown that fiducial inference has been confused with confidence interval inference. It should be added that Fisher (e.g., 8:198) used the term fiducial probability to refer to the probabilities associated with the area between fiducial limits and with the area in the tails of the fiducial distribution, i.e., the level of significance. Because Fisher employed fiducial probability in two complementary ways, latter writers have used level of significance with the same two meanings. For example, Rozeboom (17:423) speaks of a "99% level of significance" in discussing an acceptance region. By contrast, in customary usage the region of rejection, not the region of acceptance or non-rejection, is referred to in terms of level of significance (see, for example, 1:34). Given these connections, confidence interval inference with fiducial inference and level of significance with fiducial probability, it is not surprising to find level of confidence being used synonymously with level of significance--whether referring to the region of rejection or to the region of acceptance or non-rejection.

The third probable cause (from a semantic standpoint) of the confusion between level of confidence and level of significance is the bewildering set of terms associated with confidence. Examples include confidence interval, confidence level, confidence coefficient, degree of confidence, and confidence limits.

To summarize, we have suggested three possible causes for the confusion between level of significance and confidence level. These causes are: (a) the imitation of the terminology employed by well-known applied staticians, (b) the historically separate development of Fisherian and Neyman-Pearson statistics involving the deceptively similar but logically different concepts of fiducial inference and confidence interval estimation, and (c) the plethora of terms associated with the concept of confidence. Obviously, these causes are not mutually exclusive.

It would appear, then, that in most applied textbooks seldom is a logical fallacy committed in referring to level of confidence or to level of significance. Most authorities realize that, by one convention, a 5 percent level of confidence can refer to the probability of committing a Type I error. It is unhappily true that differing conventions (what we have called semantic confusion) can mislead the student, perhaps to the point of committing logical fallacies.

It is unfortunate, especially in view of the logical haze surrounding both confidence and fiducial probability, that applied staticians are so inconsistent in their choice of terminology relative to level of significance or level of confidence. Because commonly accepted conventions do have their place in the development of a discipline, we support and advocate the convention given by Mosteller and others (16) at the beginning of this paper, that is, that we restrict the magnitude of confidence level to that of the complement of the significance level. Indeed, this is the convention sanctioned by at least one of the original proponents of confidence theory (4). Furthermore, for tests of significance the experimenter is concerned with two types of error, i. e., (a) the probability of rejecting the null hypothesis when it is true and (b) the probability of failing to reject the null hypothesis when it is false. By contrast the level of confidence is a sufficient description of the potential accuracy in interval estimation. Thus although the level of significance is formally the complement of level of confidence there also is a difference in the information conveyed (see 3:429-430). Therefore, we must urge with McNemar (15) that level of confidence should be restricted to interval estimation while level of significance should refer to hypothesis testing.

REFERENCES

1. American Psychological Association, Council of Editors and Publications Board, Publication Manual of the American Psychological Association: 1967 Revision, APA, Washington, D. C. 1967, 61 pp.

2. Brownlee, K. A., Industrial Experimentation, Chemical Publishing Co., New York, 1953, 194 pp.

3. Chandler, Robert E., "The Statistical Concepts of Confidence and Significance," Psychological Bulletin, 54:429-430, September, 1957.

4. Clopper, C. J. and Pearson, E. G., "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," Biometrika, 26:404-413, 1934.

5. Dempster, A. P., "Crosscurrents in Statistics," Science, 160:661-663, May 10, 1968.

6. Edwards, Allen, E., Statistical Methods for the Behavioral Sciences, Holt, Rinehart, and Winston, New York, 1954, 542 pp.

7. English, Horace B. and English, Ava C., A Comprehensive Dictionary of Psychological and Psychoanalytical Terms, Longmans, Green and Co., 1958, 594 pp.

8. Fisher, Ronald A., The Design of Experiments (7th ed.), Hafner Publishing Company, New York, 1960, 248 pp.

9. Guilford, J. P., Fundamental Statistics in Psychology and Education (4th. ed.), McGraw-Hill Book Company, New York, 1965, 605 pp.

10. Kendall, Maurice G. and Buckland, William R., A Dictionary of Statistical Terms, Oliver and Boyd Ltd., Edinburgh, 1957, 493 pp.

11. Lehman, Paul R., Tests and Measurements in Music, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1968, 99 pp.

12. Lindquist, E. F., Statistical Analysis in Educational Research, The Riverside Press, Cambridge, Massachusetts, 1940, 266 pp.

13. Lipscomb, Edra E., "Attitudinal Change of Student Teachers," Educational Horizons, 44:99-105, Spring, 1968.

14. Lordahl, Daniel S., Modern Statistics for Behavioral Sciences, Ronald Press Co., New York, 1967, 365 pp.

15. McNemar, Quinn, Psychological Statistics (3rd. ed.), John Wiley and Sons, New York, 1962, 451 pp.

16. Mosteller, Frederick, Rourke, Robert E. F., and Thomas, George B., Probability with Statistical Applications, Addison-Wesley Publishing Co., Reading, Massachusetts, 1961, 478 pp.

17. Rozeboom, William W., "The Fallacy of the Null-Hypothesis Significance Test," Psychological Bulletin, 57:416-428, September, 1960.

18. Walker, Helen M., A review of Adam's Basic Statistical Concepts, Educational and Psychological Measurement, 16:554-557, 1956.

19. Wert, James E., Neidt, Charles O., Ahmann, J. Stanley, Statistical Methods in Educational and Psychological Research, Appleton-Century-Crafts, Inc., New York, 1954, 435 pp.

20. Wofford, J. C., "The Effects of Item Analysis Methods and Confidence Levels upon Test Validity and Cross-Validity," Journal of Educational Measurement, 5:109-114, Summer, 1968.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

COMPREHENSION OF REPEATED TIME-COMPRESSED RECORDINGS

THOMAS G. STICHT

The George Washington University [1]

ABSTRACT

Time-compressed, tape recorded messages were used to determine if listening to the message twice, in the same amount of time required to listen to the uncompressed message once, would improve listening comprehension scores of high and low aptitude men. The results indicated that for both groups of subjects, listening to the compressed messages twice improved comprehension scores over scores obtained by listening to the compressed tapes once. Comprehension of repeated time compressed messages was not improved over that obtained by a single listening to the uncompressed messages.

THE RATE of speech of a tape recorded message can be accelerated by an electromechanical process. This process involves the periodic deletion of small temporal segments of the speech, while the remaining speech segments are connected to produce connected discourse (cf., 3, for a more comlete description of this process). Through this deletion and sampling technique, the time required to present the message is decreased. Hence such speech is sometimes called time compressed speech.

If a message is time-compressed by 50 percent, it is possible to present the compressed version twice, in the same amount of time required to present the uncompressed version once. Fairbanks, and others (1) compared the comprehension of material compressed by 50 percent, but presented twice, with the comprehension of the uncompressed material requiring the same amount of time for presentation. They found a slight, but not statistically significant, improvement in comprehension with the repeated, time-compressed message. The double presentation appeared slightly more successful with men of moderate than of high mental aptitude. This suggested that the repetition procedure might prove more successful with very low aptitude men whose ability to learn by listening was being studied (4). Hence, several repetition procedures were used to determine if the learning performance of high and low aptitude men could be improved over that for a single presentation of uncompressed material.

METHOD

Materials

A selection on the use of Carbon 14 for dating relics, taken from Form lA of the Sequential Tests of Educational Progress, was used in the present study. The tape recorded listening selection was processed[2] to produce compression ratios of 0, 36, 46, 53, and 59 percents. A 20-item "fill-in-the-blank" test was prepared to evaluate listening comprehension. The test questions required the recall of factual information contained in the listening selections. Tests were scored by counting the number of correct responses.

Subjects

Eighty Army inductees served as Ss. Forty Ss had Armed Forces Qualification Test (AFQT) scores less than 30 and formed a low mental aptitude (LMA) group. The remaining forty men had AFQT scores of 87 or above and constituted a high mental aptitude (HMA) group.

Procedure

The forty men in each aptitude group were subdivided into four groups of ten. Each group of ten listened to a different pattern of repeated presentations of the listening selection. One subgroup listened to the passage compressed by 36 percent, and then the passage was repeated with 59 percent compression. A second subgroup heard the 59 percent compressed version first, and then the 36 percent compressed version. A third subgroup listened to the passage presented at 46 percent compression, followed by the same passage presented at 53 percent compression. The fourth subgroup listened to the same versions as the third group, but with the order of compression ratios reversed.

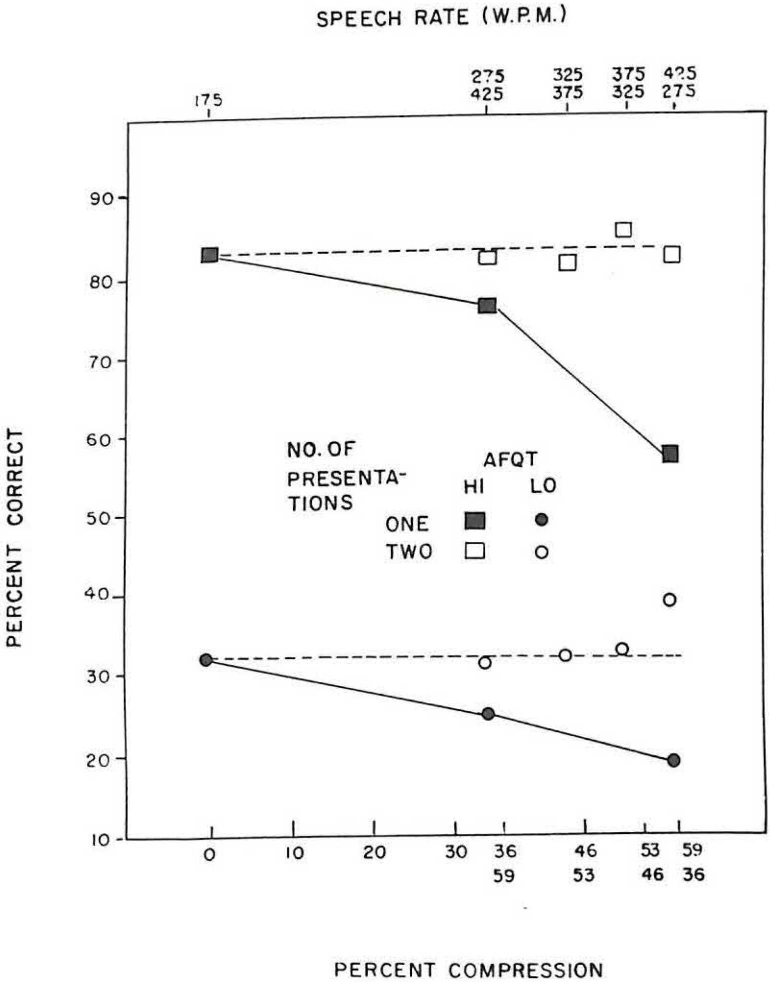


FIGURE 1

COMPREHENSION SCORES FOR HIGH AND LOW APTITUDE Ss FOR SINGLE (FILLED SYMBOLS) AND DOUBLE (UNFILLED SYMBOLS) PRESENTATION OF LISTENING SELECTION (SEE TEXT FOR EXPLANATION OF ABSCISSA).

The total listening time needed to listen to the selection once at 36 percent compression, and then again at 59 percent compression, came to 105 percent of the time required to listen to the uncompressed selection once. To listen to both the 46 and 53 percent compressed versions, the total listening time was 101 percent of the time necessary to present the uncompressed message. Listening time for the latter was 2 minutes, 35 seconds.

The recordings were presented to each S individually by means of a Wollensak tape recorder. Testing was accomplished in a sound-deadened room. Subjects listened to both versions of the repeated message before taking the comprehension test. The comprehension score was the percent of items correctly answered on the test.

RESULTS AND DISCUSSION

The results are presented as the open characters in Figure 1. In this figure, the bottom abscissa presents the compression ratios, while the top abscissa gives the corresponding speech rates in wpm. In each instance, the upper digits on the abscissa indicate the compressed version presented first. For comparative purposes, the filled symbols are presented. These symbols show how Ss similar to those of the present study scored on the same listening selection when it was presented only one time at the compression ratio indicated by the upper figure along the abscissa. These data were obtained from Ss tested in previous studies (4) who were matched as near as possible to the Ss of the present study with respect to AFQT score. Ten Ss were included in each of the mental aptitude subgroups.

The dotted lines in Figure 1 are reference lines showing the comprehension level for the single presentation of the uncompressed passage for the LMA and HMA Ss. Improvement due to listening to repeated compressed passages would be indicated by scores above the reference line. Three aspects of these data are notable. First, listening to the selection twice, with the various combination of compressed material, did not improve performance over that obtained with one presentation of the uncompressed story. This was true for both HMA and LMA groups. Second, the sequence of presentation did not differentially affect performance. The only suggestion that the sequence of presentation may have improved performance is with the LMA group and the 59 percent - 36 percent sequence. However, the score for this sequence is not significantly higher than that for the single presentation nor the presentation of the 36 percent - 59 percent sequence (Fisher exact probability test).

A final feature of interest in the data of Figure 1 is that repeating the selection with the fast word rates accompanying 46 percent and 53 percent compression resulted in better performance than would be estimated from the curves for the single presentations. The fact that there is a notable decrease in comprehension with single presentations of word rates of 325 or 375 wpm is well documented for a variety of materials (3). That the performance of the HMA and LMA groups shows improvement for the sequences having word rates of 325 and 375 wpm indicates some savings from the first presentation to the second presentation of the listening selection. But apparently the savings was not sufficient to raise the performance level above that for the single presentation of the uncompressed passage.

As mentioned earlier, Fairbanks, and others (1) reported a very slight improvement in comprehension with double presentation of compressed material over that obtained with a single presentation of uncompressed material. This improvement was more marked with men of moderate than higher aptitudes. The results of the present study, which show no improvement of peak comprehension due to repeated presentation of material, for either low or high aptitude groups, would appear to confirm the marginality of the findings of Fairbanks, and others. It should be pointed out, however, that in both the Fairbanks, and others study and the present study, technical materials were used as listening selections. Possibly the comprehension of less difficult technical or literary compressed materials might be significantly improved, with repeated presentations, over the comprehension level of the same materials presented once, in uncompressed form, in the same amount of time. But there is really no evidence to suggest this.

While the present results speak against the possibility of improving peak comprehension of listening materials through the simple repetition of materials, with men of high or low aptitudes, other research (2) suggests that the time saved by the speech compression method might be profitably used to selectively extend or review important materials. More research is needed along these lines to fully explore the educational contributions of the speech compression process for producing materials for learning by listening.

FOOTNOTES

1. The research reported in this paper was performed by HumRRO Division No. 3 (Recruit Training), Presidio of Monterey, California, under Department of Army contract with the George Washington University. The contents of this paper do not necessarily represent the official opinion of the Department of the Army.

2. The speech compression equipment was the Eltro Information Rate Changer, distributed by Gotham Audio Corp., 2 West 46th Street New York, 10036. The equipment is identified for purposes of documentation and does not imply endorsement by either Human Resources Research Office or Department of Army.

REFERENCES

1. Fairbanks, G., Guttman, N., and Miron, M. S. "Auditory Comprehension of Repeated High-Speed Messages," Journal of Speech and Hearing Disorders, 22:20-22, 1957.

2. Fairbanks, G., Guttman, N., and Miron, M. S., "Auditory Comprehension in Relation to Listening Rate and Selective Verbal Redundancy," Journal of Speech and Hearing Disorders, 22:23-32, 1957.

3. Foulke, E. and Sticht, T. G., "A Review of Research on Time Compressed Speech, in E. Foulke (Ed.), Proceedings of the Louisville Conference on Time Compressed Speech, October 19-21, 1966, Center for Rate Controlled Recordings, University of Louisville, Louisville, Kentucky, 1967.

4. Sticht, T. G., "Some Relationships of Mental Aptitude, Reading Ability and Listening Ability Using Normal and Time-compressed Speech," Journal of Communication, 18:243-258, 1968.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

THE USE OF PEER NOMINATIONS AS A PREDICTOR OF ACADEMIC SUCCESS IN COLLEGE

H. EDWIN TITUS[1]
Muskingum College

ABSTRACT

The peer nomination technique was investigated as a predictor of academic success among members of a college freshman class. The technique provided scores which were reliable and which were valid against a grade point ratio criterion. The peer nomination scores were shown to tap variance independent of SAT scores and, therefore, their applicability was suggested in academic counseling as a predictor which can be used to identify students who may face academic difficulties.

THE USE of the peer nomination technique as a predictive device in certain practical settings is becoming accepted as psychometrically sound. Military applications have involved predicting a variety of variables including: leadership qualities (2), preflight training failure (4), and promise as future officers (6). In industry the technique has been used to predict supervisory and managerial success (7, 5).

One of the intriguing aspects of the peer nomination technique is its reported success in achieving high reliability and respectable validity in groups that have been together for only relatively short periods of time. Hollander (2) achieved good results in a Naval OCS situation when the cadets had been together less than one week and had not even begun their regular training. Even better results were obtained after the groups had been together for only a few weeks. More recently Hollander (3) has reported a follow-up of his earlier work suggesting that the predictive validity of peer nomination scores hold up well against distant criteria.

The present investigation was designed to determine the efficacy of the peer nomination technique in a collegiate setting for predicting academic success among newly arrived freshmen.

METHOD

On each of three occasions, peer nomination forms calling for selections on the basis of forecasts of grades and perceptions of academic motivation were administered to 112 freshmen women and 140 freshmen men organized into nineteen dormitory units. The first occasion (time one) was at the end of "orientation week" before the subjects had attended any regular college classes. The second occasion (time two) was after 5 weeks of classes but before any grades had been officially released by the college. The third occasion (time three) was after 11 weeks of classes and after "midterm grades" had been released for advisory purposes.

The forms were administered by upper-class students who served as dormitory counselors in the various dormitory units. The counselors were fully briefed, in advance, in order to secure a good level of cooperation and to assure uniformity of conditions. They were instructed, however, to give only minimal explanation, but to assure the subjects in their units that the results would be kept confidential and used for research purposes only. The forms carried the statement:

> This is a study designed to investigate personal factors in college success. The information requested is for research purposes only and will be held in the strictest confidence. In fairness to yourself and others and to safeguard the success of this study you are asked to keep the ratings confidential. Please avoid discussing the ratings you have made with anyone.

The subjects were asked to nominate in order the members of their dormitory unit whom they expected to be the three highest and three lowest for each of the two variables: grade point ratio (GPR) and academic motivation.

Scoring followed the pattern used in other published studies (cf., 2). This procedure involved a direct weighting method to obtain peer nomination scores. Each nomination in the highest position was scored as +3, each in the next highest position was scored as +2, and each in the third highest position was scored as +1; each nomination in the lowest position was scored as -3, each in the next lowest position was scored as -2, and each in the third lowest position was scored as -1. An algebraic sum was obtained of the weighted nominations received by each subject and divided by the n of the group. A linear transformation was then computed to permit the convenience of working with a 2-digit score without an intervening decimal.

RESULTS

Table 1 presents the intercorrelations among the two peer nomination variables administered on three occasions. It is evident that the forms maintained reasonably stable results over the 11-week period and that there was substantial overlap between the two peer nomination variables. Table 2 shows the validities of the peer nomination scores and of SAT scores against earned GPR criteria. All validity coefficients were statistically significant and some were sufficiently large to have practical significance. As expected, the peer nomination scores showed substantial validity even at time one, but higher validities as a function of the length of time the groups had been

TABLE 1

INTERCORRELATIONS AMONG PEER NOMINATION SCORES FOR FORMS ADMINISTERED AT THREE DIFFERENT TIMES

Form and Time	2	3	4	5	6
1 Grade Point Ratio--time 1	.70	.54	.86	.67	.42
2 Grade Point Ratio--time 2		.72	.62	.86	.59
3 Grade Point Ratio--time 3			.48	.77	.79
4 Motivation--time 1				.65	.39
5 Motivation--time 2					.70
6 Motivation--time 3					

Note--All intercorrelations are significant at the .01 level.

TABLE 2

VALIDITY OF PEER NOMINATION SCORES AND SAT SCORES AGAINST SEMESTER GRADE POINT RATIOS

Predictors	GPR Criteria			
	1st Semester	2nd Semester	3rd Semester	4th Semester
Time 1				
Grade Point Ratio	.34	.28	.23	.22
Motivation	.28	.27	.20	.17
Time 2				
Grade Point Ratio	.39	.35	.30	.31
Motivation	.50	.45	.40	.40
Time 3				
Grade Point Ratio	.58	.42	.42	.38
Motivation	.67	.47	.42	.40
SAT				
Verbal	.37	.33	.31	.29
Math	.35	.28	.30	.24

Note--All validities are significant at the .01 level.

together. It is notable, too, that reasonably good validities were obtained against earned GPRs in later grading periods. This, of course, must be understood as largely a function of the high intercorrelation among semester GPRs (mean intercorrelation over four semesters was .67).

DISCUSSION

The primary value to be gained from a prediction of academic success among persons already selected and admitted to an academic institution is to identify as early as possible students needing special attention. In institutions which are reasonably selective, one might expect, other things being equal, that all (or at least the vast majority) of the selectees will succeed. But clearly such is rarely the case because selection is usually on the basis of ability-achievement factors, whereas it is very often the case that many other factors play a significant role in college success.

The method of peer nomination offers the possibility of utilizing students' perceptions of one another on subtle and often elusive variables (e.g., academic motivation). The data presented here suggest that the technique is sufficiently reliable and valid to be of interest. But the question remains: Does it contribute new measurement variance beyond that which was available at the time when the students were selected? Partial correlations were computed between peer nomination scores at time two[2] and first-semester grades with SAT verbal and SAT math scores held constant. The partial correlation for peer nominations of GPR with earned GPR was .34; for peer nominations of academic motivation with earned GPR it was .47. It appears that the nomination technique "tapped" variance not accounted for by the SAT scores. A multiple correlation of .56 was obtained for the two peer nomination variables (at time two) and the two SAT scores against first semester earned GPR. Beta weights for the variables were as follows: peer nominations for GPR .03, peer nominations for academic motivation .35, SAT verbal .19, and SAT math .15. Thus the peer nominations for motivation contributed the most to the prediction of the earned GPR criterion. This finding is further born out in a group-by-group[3] analysis of the data. The academic motivation variable at time two yielded validities from a low of .37 to a high of .78 (mean r = .57)[4] against earned first semester GPRs. In all but two of the dormitory units these validities were statistically significant; at time three the validities were statistically significant in all the dormitory units. Furthermore, the group-by-group analysis showed the peer nomination scores for motivation to be substantially more consistent in their validity than the SAT measures. The SAT verbal showed a range of validity coefficients from -.09 to .81 (mean r = .41). The SAT math showed a range of validities from -.29 to .55 (mean r = .32). It is likely that the relatively good results with the peer nomination motivation variable reflected the students' understanding of the implications of academic motivation. That is, as something separate from ability and achievement--an enthusiasm, a seriousness of purpose, an air of congruence with the aims and purposes of an academic institution. These data do not, of course, identify what their conception of academic motivation was but only that was relatively independent of the SAT kind of variance (this interpretation was further strengthened by results of factor analyses not reported here).

Rating one's classmates is difficult and, for some, a trying task. It was suspected that attitude toward the process might materially affect the accuracy of predictions and that there might be sex differences even overriding this additudinal factor. With regard to the latter consideration, there were no significant differences between men and women in either the reliability or validity of their peer nominations. It is perhaps notable that the women at time one and time two expressed slightly more resistance to the process of making interpersonal judgments than did the men. A very few subjects of both sexes declined to make any nominations at all, but the vast majority even of those who were reluctant and/or skeptical performed as requested.

A section of the cover sheet accompanying the peer nomination forms offered the subjects an opportunity to make comments. Most of the subjects made no comments at all or made comments that were judged to be neutral. However, 37 percent of the subjects made comments at time one that were skeptical or negative compared with only 8 percent at time two and 3 percent at time three. Examination of these data on a group-by-group basis showed that the groups which produced the highest frequencies of negative reactions were indistinguishable from the remaining groups with respect to reliability and validity of the peer nomination scores.

SUMMARY AND CONCLUSION

The peer nomination technique was applied at three times during their first semester at college to 252 freshmen organized into nineteen dormitory units. Scores based on their nominations concerning GPR and academic motivation showed high reliability and promising validity. Since the peer nominations were seen to be measuring variance independent of the ability-achievement factors usually considered in academic predictions, it appears that the technique can provide unique information and perhaps can contribute to the early detection of students who are in need of special academic counseling.

FOOTNOTES

1. The author wishes to express his appreciation to Miss Patricia Siller who provided valuable assistance in the pilot work preceding this study.

2. Time two was chosen because it appeared to offer the best compromise for practical applications between the need to have information early and the need to have it as accurate as possible.

3. All analyses based on individual dormitory units were limited to the twelve groups with $n \geq 12$.

4. Mean r's reported in this paper are based on Fisher's (1) $\underline{z}$ transformation.

REFERENCES

1. Fisher, R. A., "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," Biometrika, 10:507-521, 1915.

2. Hollander, E. P., "The Reliability of Peer Nominations Under Various Conditions of Administration," Journal of Applied Psychology, 41:85-90, 1957.

3. Hollander, E. P., "Validity of Peer Nominations in Predicting a Distant Performance Criterion," Journal of Appiled Psychology, 49:434-438, 1965.

4. O'Connor, W. F., and Berkshire, J. R., "Comparison of the Pre-flight OLQ Grade and the Leadership Peer Rating as Predictors of Training Failure," U. S. Naval School of Aviation Medicine, Special Report No. 58-16.

5. Roadman, H. E., "An Industrial Use of Peer Ratings," Journal of Applied Psychology, 48:211-214, 1964.

6. Suci, G. J., Vallance, T. R., and Glickman, A. S., "A Study of the Effects of 'Likingness' and Level of Objectivity on Peer Rating Reliabilities," Educational and Psychological Measurement, 16:147-152, 1956.

7. Weitz, J., "Selecting Supervisors With Peer Ratings," Personnel Psychology, 11:25-35, 1958.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

THE DEVELOPMENT OF PRODUCTIVE THINKING SKILLS IN FIFTH-GRADE CHILDREN

JAMES L. WARDROP
University of Illinois

WILLIAM L. GOODWIN
Bucknell University

HERBERT J. KLAUSMEIER
University of Wisconsin

ROBERT M. OLTON
University of California, Berkeley

MARTIN V. COVINGTON
University of California, Berkeley

RICHARD S. CRUTCHFIELD
University of California, Berkeley

TECKLA RONDA
Wisconsin Public Schools, Racine

ABSTRACT

A controlled experiment was conducted in forty-four fifth-grade classrooms to investigate the extent to which creativity and problem-solving skills of children could be nurtured through a series of self-instructional programmed lessons. Also studied was the relationship of such skills to (1) learner characteristics (IQ and sex) and (2) classroom "environment" as it may have facilitated creative thinking. Significant differences were found favoring the experimental group (which received the programmed lessons). Improvement in productive thinking skills was found for both boys and girls of both higher and lower IQ. Greater gains were found in classrooms providing relatively little support and encouragement for productive thinking. Also the performance of girls on the productive thinking measures exceeded that of boys and there was a strong positive relationship to IQ.

CREATIVITY in school children--particularly gifted children--has been studied extensively (8, 10). This research has led to the organization of programs for the gifted which include education for creativity (7, 9). There is now increasing interest in educational programs which seek to develop the creative potential of a much wider segment of the school population. One set of instructional materials which attempt to do this is The Productive Thinking Program, Series One: General Problem Solving (3). Developed by researchers at the University of California at Berkeley, this program consists of a set of programmed lessons intended to develop and strengthen productive thinking skills in fifth- and sixth-grade students. In previous studies, these materials produced a substantial increase in the creative thinking performance of students at many different levels of intellectual ability (2, 4).

The major purpose of this study was to investigate the extent to which the creativity and problem-solving skills of fifth-grade children could be improved through use of such a set of programmed lessons. In this context, several specific aims were incorporated:

(1) To investigate further the extent to which productive thinking (i. e., creative thinking and problem solving) abilities can be taught to elementary school children independent of any specific subject field.

(2) To discover whether some of these skills or behaviors are more readily taught than others. For example, one aspect of problem solving is the asking of appropriate questions. Can students be taught this behavior more (or less) successfully than they can some other skill, such as generating a set of possible solutions to a problem?

(3) To investigate the relationship between certain learner characteristics (intelligence and sex) and performance on a broad diversity of productive thinking measures. It is generally reported that girls perform slightly better than boys on most measures of academic achievement. Is this also true of measures of productive thinking? It has also been reported (14) that intelligence and "creativity" are relatively independent. Does this independence persist throughout several different aspects of productive thinking and on several different types of problems? In addition, is there a relation between intelligence or sex and the ability to learn productive thinking skills?

(4) To investigate the extent to which overall classroom "Environment" facilitates productive thinking in students, and the extent to which students

in different types of classroom environments profit from instruction in productive thinking. The investigators felt that the kind of classroom atmosphere (provided primarily by the teacher in her conduct of class discussions) might influence the kind and amount of productive thinking on the part of students in that classroom. Students in a classroom where the teacher encourages a free and open exchange of ideas might tend to exhibit more of what is here considered productive thinking behavior than would students in a classroom where the teacher maintains close control over the ideas which may be pursued in discussions. To what extent does merely providing an encouraging environment promote creative thinking? Are instructional materials more (or less) effective in such environments?

(5) To determine the effectiveness of Series One of The Productive Thinking Program when these materials are used as a self-contained instructional package. In this experiment, teacher participation was purposely held to a minimum, as might be the case when a particular student or group is assigned an independent reading program, to be done with little supervision by the teacher.

METHODS

Subjects and Conditions

Of the more than sixty fifth-grade classrooms in the Racine, Wisconsin, Unified School District No. 1, a number were involved in other studies and were not included in this one. Of the forty-seven available classrooms, forty-four were employed. All students in these forty-four classes participated in the experiment, but for simplification analyses were based upon a sample of eight males and eight females selected from each class. Selection was accomplished by rank-ordering all male students from a given classroom on the basis of IQ, and then randomly selecting one student from each eighth of the distribution. Eight females were then selected from the same class in the same way. This procedure was continued for all forty-four classes, so that the entire analysis involved a total of 704 students. (Selection was made after the experiment had been completed but before any scoring of tests was begun.)

This large sample reflected, in proper proportion, the wide variety of abilities and backgrounds characteristic of students in an urban school system (a city of about 100,000 population) and hence provided a thoroughly representative base for analyzing both overall and internal effects. The size of the sample also permitted a very stable estimate of error, even with the use of classroom means (rather than individual student scores) as the unit of analysis.

Several weeks before the experiment began, the teaching supervisor of the district observed each class while the teacher and students were having an extended discussion about a standardized hypothetical event in social studies class, and recorded the extent to which the overall classroom atmosphere (including the teacher) seemed to provide an environment that fostered creative thinking on the part of the students. The supervisor reported that she felt these ratings were representative of the normal state of affairs that existed in the various classrooms. The final, single Environment rating assigned to each class was a composite score derived from the several objective ratings the supervisor made. Those ratings were designed to indicate the extent to which the teacher allowed free and open discussion, the degree to which the teacher avoided making value judgments about unusual ideas offered by students during class discussion, and other characteristics related to the extent to which the classroom environment could be thought of as facilitating or inhibiting the kinds of behavior considered indicative of productive thinking. The twenty-two classrooms with the highest ratings were designed as Facilitative Environment classrooms; likewise, those twenty-two classrooms whose composite rating was below the median were classified as Nonfacilitative Environment groups.

Within each of these Environment blocks, half the classrooms were randomly assigned to the Treatment Group (which would use The Productive Thinking Program) and the remaining half constituted the Control Group.

Instructional Materials

The Productive Thinking Program (3) is a set of instructional materials which undertakes to develop the student's ability to use important skills and strategies of thinking and problem solving, seeks to enhance his understanding of the processes of his own thinking, and aims to promote favorable attitudes and motivations toward activities which involve the use of the mind. Series One of this Program, which deals with General Problem Solving, served as the instructional materials in this study.

The skills taught by Series One of The Productive Thinking Program are those thought to be essential for the kind of complex and original thinking in which an individual strives to produce new ideas, to attain fresh insights and understandings, and to achieve his own solutions to problems. The program also shows the student how to proceed in a systematic way when attacking a problem, how to clarify a problem, and how to attend to relevant facts and conditions of the problem in evaluating his ideas. In addition, it seeks to promote those motivations and attitudes required for effective performance in productive thinking tasks.

Series One of The Productive Thinking Program consists of sixteen programmed lessons, each of which is an individual booklet about 40 pages long. The lessons have a cartoon-text format which serves to heighten interest and at the same time permits clarity and emphasis in the presentation of the material. The lessons are written primarily for fifth- and sixth-grade students and can be used by students with a wide range of intelligence within these grades. Each lesson is self-administering, allowing the student to work at his own pace; virtually all students complete a lesson within 50 minutes. A detailed description of the materials may be found in Olton and others (12).

The 16-lesson series requires a total of 16 classroom hours of instruction time. In the present study, the lessons were administered one per day for 4 days of each week; the remaining school day in each week was used for makeup sessions. Generally, each student worked individually at his own pace and had an assignment to which he turned when he had completed the day's lesson.

Since it was the intention of this study that the teacher's role in the use of these materials be held to a minimum, the teachers were instructed only to distribute the lessons and answer procedural questions. They were not to conduct class discussions based on the materials or otherwise participate in the way they were used. (One exception to this policy was permitted in classes with a number of poor readers. In these classes, teachers often read portions of the lessons aloud with small groups, or otherwise assisted students who had substantial reading problems.)

Measuring Instruments and Scoring Procedures

Most of the measuring instruments used in this study were developed by the Creative Thinking Project at Berkeley and differ from traditional educational tests in both content and structure. 2 A discussion of their underlying rationale may be found in Covington (1).

Briefly, the instruments are based on the proposition that creative thinking can be conceptualized as an intricate and highly complex problem solving process which requires the individual to use and to coordinate a number of different cognitive skills as he seeks to reach his goal. This complex problem-solving process involves: (1) the use of a number and variety of cognitive skills; (2) the effective management and deployment of these cognitive skills by what might be called a "master" thinking skill; and (3) a high degree of involvement on the part of the thinker--the individual is highly engrossed with the creative task and finds it meaningful and engaging.

Hence the measuring instruments used in this study attempt to present the student with thinking tasks which he will find meaningful and engaging, and which require him to use and to coordinate a number of different cognitive skills. These tasks do not involve the use of specialized knowledge or technical skills, nor do they require the mastery of any particular curriculum material.

A complete list and classification of the tests of productive thinking used in this study is presented in Table 1. In this classification, the terms "convergent" and "divergent" thinking are used in the traditional way to denote thinking which, on the one hand, involves processing information so as to "converge" upon a single solution to a problem or, on the other hand, involves the use of information as a point of departure from which one's thinking "diverges" into a variety of new mental products. "Complex Extended Problems" in the classification are those which feature a gradual development of the problem and require the student to use a number of different cognitive skills as he proceeds. The student responds at several points and is presented with feedback in the form of more information about the problem after each response opportunity. Those measures labeled "Brief Problems" require the use of relatively few cognitive strategies and do not employ response-feedback sequences; instead, they present the entire problem to the student at once, and he then makes his response(s).

All the problems are further classified as verbal or nonverbal, and whether they were used as pre- or posttests or as internal tests in this study. The name of each test and the indices of thinking proficiency on which it was scored are also listed.

TABLE 1

CLASSIFICATION OF THE TESTS OF PRODUCTIVE THINKING USED IN THIS STUDY

Name of test	Use	Performance indicators
	1. Complex Extended Problems with an Emphasis on Convergent Thinking (all verbal)	
Pit problem	pretest	number of ideas; quality of ideas (two indices); achievement of solution
Jewel problem (short form)	pretest	number of ideas, quality of ideas (two indices); achievement of solution
Rare Coins problem	internal test	achievement of solution
Jewel problem (long form)	posttest	number of ideas; quality of ideas (two indices); achievement of solution
Black House problem	posttest	number of ideas; quality of ideas (two indices); achievement of solution; number of odd facts or discrepancies in information that are noted
X-ray problem	posttest	number of ideas; quality of ideas (two indices); achievement of solution; number of relevant questions asked or preparatory steps taken before work on problem was begun

2. Complex Extended Problems with an Emphasis on Divergent Thinking (all verbal)

Elevator problem	pretest	number of ideas
The Deep Sea Dive	internal test	number of odd or puzzling facts within the problem that are noted; achievement of solution
The Ancient City	posttest	number of ideas; quality of ideas (two indices); number of relevant questions asked or preparatory steps taken before work on the problem was begun
A Visit to Karam	posttest	number of puzzling facts which are noted; intellectual persistence

3. Brief Problems with an Emphasis on Convergent Thinking (all verbal)

Ingenuity Items (Form A)	pretest	achievement of solution (number of solutions achieved: three possible)
Ingenuity Items (Form B)	posttest	achievement of solution (number of solutions achieved: ten possible)

4. Brief Problems with an Emphasis on Divergent Thinking

Squares	pretest nonverbal	number of ideas; quality of ideas (two indices)
Truck Improvement	pretest verbal	number of ideas; quality of ideas (two indices)
Suppose People Could Read Minds	pretest verbal	number of ideas; number of inventions suggested
A Summer Carnival (Form A)	pretest verbal	quality of ideas (two indices); number of available elements used when creating a product
Circles	posttest nonverbal	number of ideas; quality of ideas (two indices)
Dog Improvement	posttest verbal	number of ideas; quality of ideas (two indices)
Suppose There was an Intelligence Pill	posttest verbal	number of ideas; number of inventions suggested
A Summer Carnival (Form B)	posttest verbal	quality of ideas (two indices); number of available elements used when creating a product

5. Test of the Master Thinking Skill (verbal)

The Missing Statue	posttest	total number of "best" or "second best" alternatives to which the student gives a rank of one; number of occasions on which the student chooses the "give up" alternative

Performance Indicators and Scoring

On each of the various tests, several measures of performance were obtained, the particular measures employed depending on the nature of the problem. For an example which illustrates the general nature of a complex extended problem and presents a detailed discussion of the rationale and procedures by which performance was scored, see Olton and others (12).

A number of indices of thinking proficiency can be applied to performance on such problems. The following performance indicators are typical of those which were used: whether or not the student achieved the principal solution to the problem; the number of odd facts or discrepancies in the problem which he noticed; the number of ideas he wrote as possible solutions to the problem (regardless of the quality or adequacy of these ideas); and, finally, two indices of the quality of the ideas he produced.

Quality ratings were made by a team of trained scorers who first assigned a separate rating to each of a student's ideas in a problem, and then recorded both the number of "high quality" ideas and the quality rating of the "best" idea which the student produced.

In addition to measuring the student's performance on problem-solving tasks, an attempt was made to probe his understanding of the nature of problem solving by presenting him with three open-ended sentences which he was asked to complete. The sentences were: "Making mistakes in working on a problem is ________"; "Wild or silly ideas are ________"; and, "In order to solve problems you must ________." The student's answers rated by scorers on a 3-point scale, indicating the degree of insight or understanding which was shown by the answer. Each of the three sentences was scored as a separate variable and is referred to as Sentence I, II, or III, respectively, in the analysis.

Experimental Design and Statistical Analysis

The basic analyses for this experiment were done with a 2 x 2 x 2 factorial design. The factors represented were: Instruction (Treatment vs. Control); Environment (Facilitative vs. Nonfacilitative); and Sex. There were eleven classrooms within each of the cells of this design, and standard analysis of variance procedures were used for the analyses.

Because the investigators also wished to look at the relationships between IQ and the Treatment and Environment factors, a secondary analysis was performed, using a 3 x 2 x 2 factorial design. Three levels of IQ were represented in this design--high (above 115), middle (101 through 115), and low (100 and below). The other two factors were Treatment-Control and Facilitative-Nonfacilitative Environment. After all subjects had been blocked into three groups on the basis of IQ, a random sample of twenty was selected at each IQ level within each Treatment x Environment combination in the design. Thus, a total of 240 subjects was used for this secondary analysis. (In order to obtain such a sample, it was necessary to ignore classifications according to sex and classroom units involved in the experiment.)

For both the main analysis (the 2 x 2 x 2 factorial design) and the secondary analysis (the 3 x 2 x 2 factorial design involving IQ), separate analyses of variance were performed on each of the pretest, internal test, and posttest measures indicated in Table 1. Similar analyses were performed on each of five stratification variables--socioeconomic status, overall mean score on the Stanford Achievement Battery, IQ, performance on the Social Studies part of the STEP test (Sequential Test of Educational Progress) when given as a pretest (Form 4A), and scores on this test when given as a posttest (using Form 4B).

In the main design, the classroom mean score for each variable was the unit of analysis, while in the second design, the analysis was performed using scores of individual students as the basic unit.

RESULTS[3]

Analyses of variance of the stratification variables yielded no significant treatment-control group differences. Sex differences were found for Stanford Achievement Test (F=7.81; df = 1, 80; $p < .01$) and the Kuhlmann-Anderson IQ scores (F=8.81; df = 1, 80; $p < .01$). In both cases, females scored higher than males (0.4 years on the Stanford and 4.55 points on the IQ test). Significant Environment effects were found on all stratification variables (Table 2); the means are presented in Table 3.

TABLE 2

MEAN SQUARES AND F RATIOS FOR SIGNIFICANT ENVIRONMENT EFFECTS: STRATIFICATION VARIABLES (1 AND 80 DEGREES OF FREEDOM)

Variable	Mean Square	F
Socioeconomic status	24.04	11.58**
Stanford achievement	2.09	5.64*
Kuhlmann-Anderson IQ	484.02	9.37**
STEP pretest	47.39	4.55*
STEP posttest	42.62	4.30*

*$p < .05$
**$p < .01$

TABLE 3

MEAN LEVEL OF PERFORMANCE OF STUDENTS IN FACILITATIVE AND NONFACILITATIVE ENVIRONMENTS: STRATIFICATION VARIABLES

	Environment	
Variable	Facilitative	Nonfacilitative
Socioeconomic status	2.45	3.50
Stanford achievement	5.25	4.93
Kuhlmann-Anderson IQ	109.74	105.05
STEP pretest	27.18	25.71
STEP posttest	26.18	24.79

Pretest Results

In the twenty-one analyses of pretest scores, only one treatment effect was significant, for the measure of Achievement of Solution on the Jewel Problem (F = 5.46; df = 1, 80; $p < .05$). This difference reflected the superiority of students in the treatment group (8 percent achieving the correct solution as opposed to only 4 percent in the control group). It seems likely that this significant F is simply an artifact of the number of analyses carried out. Table 4 presents the mean squares and F ratios for the significant Environment effects on the pretest measures, with the corresponding means in Table 5. Tables 6 and 7 present the mean squares, F ratios, and means for those variables on which significant sex differences were found. The means in Table 5 indicate the consistently superior performance of students in the facilitative environments, and those in Table 7 show that males scored higher on three measures (all related to the same problem), while females scored higher on the other five.

TABLE 4

MEAN SQUARES AND F RATIOS FOR SIGNIFICANT ENVIRONMENT EFFECTS: PRETEST (1 AND 80 DEGREES OF FREEDOM)

Variables	Mean Square	F
Squares problem		
Number of ideas	23.66	4.08*
Number high quality ideas	5.73	8.90**
Quality rating, best idea	0.32	4.70*
Truck improvement problem		
Number high quality ideas	2.04	5.37*
Quality rating, best idea	0.27	5.56*

* $p < .05$
** $p < .01$

TABLE 5

MEANS FOR STUDENTS IN FACILITATIVE AND NONFACILITATIVE ENVIRONMENTS FOR SIGNIFICANT PRETEST VARIABLES

Variable	Environment	
	Facilitative	Nonfacilitative
Squares problem		
Number of ideas	8.97	7.93
Number high quality ideas	2.18	1.67
Quality rating, best idea	1.71	1.59
Truck improvement problem		
Number high quality ideas	1.58	1.27
Quality rating, best idea	1.69	1.58

Internal and Posttest Results

Table 8 presents the mean squares and F ratios for the significant Treatment effects, thirteen of which are statistically significant (as opposed to one such difference that was significant on the pretest measures). The means for these measures are presented in Table 9. Those measures involving the X-ray problem require special comment, since it was the only one in the entire posttest battery on which the control group consistently outperformed the treatment group by a statistically significant margin. This reversal of the general pattern of results appears to be an artifact of this particular problem, possibly because the task was so difficult (only 15 percent of all students ever achieved either of the two "best" solutions) that an outstanding performance by relatively few students in the control classes could produce significant differences favoring that group.

Analyses of variance of the three internal test measures and thirty-seven posttest variables resulted in one significant F ratio for the Environment effect, on the Ingenuity Test ($F = 4.89$; $df = 1, 80$; $p < .05$). This is a considerable contrast to the pretest results in which five significant differences (out of twenty-one analyses) were found, all of which indicated the superiority of students in facilitative environments. The notable lack of posttest effects may partly be accounted for by the presence of several significant Treatment x Environment interactions (see below), all of which indicate that treatment was more effective with students in nonfacilitative environments, thus reducing the initial differences between facilitative and nonfacilitative groups. Several additional Treatment x Environment interactions, though they failed to reach significance, also showed this pattern.

TABLE 6

MEAN SQUARES AND F RATIOS FOR SIGNIFICANT SEX EFFECTS: PRETEST (1 AND 80 DEGREES OF FREEDOM)

Variable	Mean Square	F
Pit Problem		
Achievement of solution	0.41	11.59**
Number of ideas	2.21	4.27*
Number high quality ideas	1.18	9.31**
Quality rating, best idea	2.16	13.07**
Mind-reading problem		
Number of ideas	15.54	8.38**
Summer carnival		
Number high quality ideas	0.47	5.67*
Quality rating, best idea	0.47	5.10*
Ability to synthesize	5.77	19.11**

* $p < .05$
** $p < .01$

TABLE 7

MEANS FOR MALES AND FEMALES FOR SIGNIFICANT SEX EFFECTS: PRETEST

Variable	Males	Females
Pit problem		
Achievement of solution	0.49	0.35
Number of ideas	3.47	3.79
Number high quality ideas	1.19	0.96
Quality rating, best idea	3.17	2.86
Mind-reading problem		
Number of ideas	3.52	4.36
Summer carnival		
Number high quality ideas	0.33	0.48
Quality rating, best idea	1.03	1.17
Ability to synthesize	1.84	2.36

TABLE 8

MEAN SQUARES AND F RATIOS FOR SIGNIFICANT TREATMENT EFFECTS: INTERNAL AND POST-TESTS (1 df FOR EFFECTS)

Variable	Mean Square	F
Deep sea dive problem (internal)[a]		
Sensitivity to incongruous information	2.25	7.41**
Jewel problem:		
Number of ideas[b]	13.46	6.59*
House problem[a]		
Achievement of solution	0.23	6.47*
Number high quality ideas	2.05	16.99**
Quality rating, best idea	1.84	5.98*
Sensitivity to incogruous information	1.50	15.65**
X-ray problem[a]		
Achievement of solution	0.16	6.20*
Quality rating, best idea	0.99	5.22*
Ancient city problem:		
Number of ideas[b]	22.66	10.97**
Visit to Karam:		
Persistence[b]	2.74	4.15*
Intelligence pill:		
Inventiveness[a]	0.38	4.33*
Statue problem:		
Effectiveness of Master Thinking Skill[a]	1.97	5.06*
Understanding thinking:		
Sentence I[b]	0.68	6.79*

*$p < .05$
**$p < .01$
[a] 80 df for error
[b] 36 df for error

Each of the significant sex differences reported in Table 10 (twelve in forty analyses) reflects the superior performance of females. Seven of the twelve differences occurred on variables dealing with the production of ideas, three differences reflected the greater sensitivity of females to discrepant or puzzling information embedded in the tests, and the remaining two differences occurred on a variety of measures. With respect to the verbal fluency measures, the data are consistent with the numerous findings that girls are verbally more proficient than boys.

In the analyses of all these measures--pretest, internal test, and posttest--there was a total of seven significant interactions, all in the internal and posttests. Two of these, three-way interactions of Treatment x Environment x Sex, were uninterpretable. The other five were Treatment x Environment interactions for which the means are presented in Table 11. The interactions for these five measures all have the same pattern: within the treatment groups those students in nonfacilitative environments performed better than students in facilitative environments, while for the control group, students in the facilitative environments performed better than those in the nonfacilitative environments. Thus, it would seem that one important effect of the training materials was to enable students from less favorable classroom climates to overcome the detrimental effects that such environments would otherwise have produced on their performance. This may help explain the fact (noted earlier) that classroom environment was a significant source of variance on five out of twenty-one pretest measures, but on only one out of forty posttest measures.

TABLE 9

MEANS FOR TREATMENT AND CONTROL GROUPS FOR SIGNIFICANT POSTTEST EFFECTS

Variable	Group: Treatment	Group: Control
Deep sea dive problem		
Sensitivity to incongruous information	1.63	1.29
Jewel problem:		
Number of ideas	6.67	5.58
House problem		
Achievement of solution	0.70	0.59
Number high quality ideas	1.32	1.01
Quality rating, best idea	4.23	3.94
Sensitivity to incongruous information	0.66	0.40
X-ray problem		
Achievement of solution	0.11	0.19
Quality rating, best idea	2.48	2.69
Ancient city:		
Number of ideas	5.22	4.06
Visit to Karam:		
Persistence	2.70	2.22
Intelligence pill:		
Inventiveness	0.08	0.03
Statue problem:		
Effectiveness of Master Thinking Skill	3.73	3.32
Understanding thinking:		
Sentence I	0.78	0.53

TABLE 10

MEAN SQUARES AND F RATIOS FOR SIGNIFICANT SEX EFFECTS: INTERNAL AND POSTTESTS (1 df FOR EFFECTS)

Variable	Mean Square	F
Deep sea dive problem (internal)[a]		
Sensitivity to incongruous information	1.25	4.11*
House problem[a]		
Number high quality ideas	0.57	4.71*
Sensitivity to incongruous information	0.67	7.00**
X-ray problem:		
Sensitivity to incongruous information[a]	1.70	9.59**
Circles:		
Number of ideas[a]	55.63	9.37**
Dog improvement[a]		
Number of ideas	57.05	14.13**
Number high quality ideas	1.52	4.55*
Intelligence pill:		
Number of ideas[a]	6.61	9.99**
Summer carnival[a]		
Number high quality ideas	0.98	16.97**
Quality rating, best idea	1.70	20.49**
Ability to synthesize	4.99	15.41**
Understanding thinking:		
Sentence I[b]	0.81	8.12**

*$p < .05$
**$p < .01$
[a] 80 df for error
[b] 36 df for error

Effects of IQ

The purpose of the second analysis was to determine the effect of IQ on performance, both as a main effect and in its interactions with Treatment or Environment. Since the main effects of the Treatment and Environment factors have already been discussed, they will not be considered in connection with this analysis.

With the exception of socioeconomic status, performance on each of the stratification variables was significantly related to IQ ($p < .01$). The mean differences for these variables are all in the predicted direction--high IQ students perform better on the Stanford and STEP tests than do middle IQ students, who in turn perform better than those with low IQ.

With respect to the pretest measures, IQ represented a significant source of variance for seventeen of the twenty-one measures (fifteen of these beyond the .01 level). For all of these problems, the high IQ group outperformed the others, and the middle IQ group, with one small reversal, was consistently superior to the low IQ group. Significant differences were found on twenty-seven of the thirty-eight posttest measures, nineteen of these beyond the .01 level. (IQ effects on the three internal test measures were not analyzed.) In all cases, the high and middle IQ groups performed substantially better than the low IQ group. However, performance of the high and middle IQ groups on these posttest measures was much more alike (relative to the low IQ group) than it had been in the pretest. The absence of any systematic IQ x Treatment interaction (see below) suggests that this change is not of sufficient magnitude to be considered an important result of exposure to the program. As in the pretest, there was a general rank-order relation between IQ and performance on most of the posttest measures.

A total of nine significant interactions occurred in this second analysis, four involving pretest measures and the other five involving the posttest. No consistent pattern appears to be present in these interactions.

DISCUSSION

Effects of Treatment

Clearly, use of the program led to an increase in the level of productive thinking in this large and representative group of fifth-grade students. Mean performance of the group which received training in productive thinking skills surpassed that of the control group on thirty of the forty internal and posttest measures. Of the thirteen significant differences, eleven favored the treatment group. Moreover, this superiority was found at all levels of IQ, and was evident on a broad variety of the measures.

However, there was a lack of treatment effects on nearly all performance measures for the brief problems, most of which emphasized divergent thinking. These brief tests, such as those developed by Torrance, are structured so as to elicit a large number of rather short, discrete responses on relatively uncomplicated tasks. One possible explanation for the lack of such effects focuses on the question of the relevance of the training materials to such skills. If it is assumed that the brief tests require less complex cognitive processes for adequate performance than those processes taught for in the complex problem settings of Series One, then what is learned in Series One is only marginally relevant to performance on these brief problems. Thus, the treatment group might not show any marked superiority on such tasks. Another possibility which must be considered is that most children of this age, whether trained or not, can demonstrate these relatively simple cognitive skills on demand (under proper conditions of short-term motivation), even though they may rarely make use of them in their regular school work. If this is true, training provided by the program would produce superior performance on complex thinking tasks, where the use of such simple cognitive skills (and a positive set) is not sufficient for successful performance, but no obvious benefit from training would be apparent on the brief tests.

Effects of Environment

The five significant Environment effects on the pretest measures all resulted from the fact that classes with facilitative environments performed better than nonfacilitative classes on the two brief divergent thinking problems developed by Torrance. On the posttest and internal test measures, there was only one significant Environment effect out of forty comparisons. This virtual absence of post-training Environment effects seems to occur in large part because the treatment was particularly effective with students in nonfacilitative environments. Nearly a dozen Treatment x Environment interactions (five of which are significant and are presented in detail in Table 11) support this conclusion. It should be noted that the special benefits of training for students in nonfacilitative classes amounted to more than simply a change of "set," since some of the significant Treatment x Environment interactions involved the kind of complex problems in which no main effect of Environment was found. Thus it appears that an important effect of the training materials was to improve both the cognitive skills and the general attitudes of students from the less favorable classroom environments.

Effects of Sex

Significant sex differences occurred on twenty-one of the total of sixty-one pretest, posttest, and internal test measures used in this study. With the exception of three measures in the pretest Pit Problem, all significant sex effects resulted from the superior performance of females. This female superiority occurred on all types of problems (both verbal and nonverbal, convergent and divergent), and on all performance indicators except measures of intellectual persistence and of the achievement of solution to problems. This finding of generally superior performance by females is consistent with other findings (e.g., 10) concerning the verbal performance of students of this age.

Effects of IQ

The general level of intellectual ability had a clear and significant effect upon performance for most of the measures used in this study. In most cases, the predictable rank-order relation between IQ and performance was found, with the High IQ group obtaining scores superior to those obtained by the Middle IQ group, which were in turn superior to those obtained by the Low group. The effect of IQ was significant on all types of performance measures, and on all types of problems except the special test of the Master Thinking Skill.

Large and significant IQ effects were found on all of the brief divergent thinking problems, and this effect was especially marked on the four Torrance tests. These effects were often large in an absolute sense as well as significant statistically. This set of findings suggests that early reports of a "virtual lack of relationship between measures of creative thinking and IQ" (14) are perhaps something of an overstatement and that such findings may have resulted from sampling restrictions. Since all IQ analyses

TABLE 11

MEANS FOR SIGNIFICANT TREATMENT x ENVIRONMENT INTERACTIONS (1 df FOR INTERACTION)

Variable	df for error		Environment	
			Facilitative	Nonfacilitative
Rare coins problem (internal test):	80			
Achievement of solution		Treatment	.23	.22
		Control	.28	.12
Jewel problem (posttest):	36			
Number of ideas		Treatment	6.25	7.21
		Control	6.03	5.21
Dog problem (posttest):	80			
Number of ideas		Treatment	8.10	8.46
		Control	8.28	6.86
Number high quality ideas		Treatment	1.20	1.31
		Control	1.34	.92
House problem (posttest):	80			
Sensitivity to incogruous information		Treatment	.54	.79
		Control	.48	.33

in the present study are based upon the carefully selected subsample of 240 students (described earlier) which should be representative of the entire distribution of ability in a total of forty-four classrooms, the highly significant IQ effects obtained cannot easily be dismissed as an artifact of the particular sample used.

Indeed, it seems reasonable to assume that there is a moderate correlation between IQ and creative thinking when the entire distribution of intellectual ability is considered. Most studies which have reported the absence of such a correlation (e.g., 5, 15) involved highly selective, nonrepresentative samples, while studies which reported a positive relation between IQ and creative thinking (e.g., 6, 13) have used far more representative samples.

In contrast to the effect of IQ on test performance, there was no consistent pattern of interactions between IQ and Treatment; that is, the increments in thinking and problem-solving performance produced by the training program were approximately equal for all three levels of IQ. This finding agrees with the results of the earlier Covington and Crutchfield study (2), and also with the more recent study by Olton and Crutchfield (11).

SUMMARY

The major purpose of this study was to investigate the extent to which increments in the thinking and problem-solving performance of fifth-grade students could be produced by the use of self-instructional programmed lessons designed to teach skills and strategies of creative thinking independent of any specific subject field. The study also investigated the relationship between productive thinking abilities and certain characteristics of the learner (IQ and sex), and the relationship between level of productive thinking performance and the extent to which overall classroom "environment" was judged to facilitate creative thinking.

Results from forty-four fifth-grade classes showed that the training materials produced statistically significant increments in thinking and problem-solving performance on a wide variety of productive thinking measures. These instructional benefits occurred for virtually all types of students (regardless of sex or level of IQ), and were especially marked for students in classrooms having environments which were judged to provide relatively little support and encouragement for the development of productive thinking. These effects were obtained when the materials were used as an entirely self-contained, self-instructional program; considerably greater educational benefits could be expected under conditions where materials are reinforced by active teacher participation.

Quite apart from the effects of the instructional materials, performance on the productive thinking measures used in this study was significantly related to sex (girls generally scoring higher than boys) and showed a strong and positive relation to IQ.

FOOTNOTES

1. The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare, under the provisions of the Cooperative Research Program (Center No. C-03/Contract OE 5-10-154, Wisconsin Research and Development Center for Cognitive Learning) with additional support of a grant from the Carnegie Corporation of New York to the Creative Thinking Project, Institute of Personality Assessment and Research, University of California, Berkeley.

2. Four tests in the battery (Circle, Squares, Truck Improvement, and Dog Improvement) were developed by Dr. E. Paul Torrance while at the University of Minnesota. His permission to use those tests as part of this study is gratefully acknowledged.

3. A detailed presentation of the results of this experiment, including tables reporting means, standard deviation, and results of analyses of all variables will be found in Olton and others (12). Further analyses are currently being conducted by Mr. James Bavry of the Wisconsin Research and Development Center for Cognitive Learning, Madison.

REFERENCES

1. Covington, M. V., "New Directions in the Appraisal of Creative Thinking Potential," Journal of Educational Measurement, 1969, in press.

2. Covington, M. V., and Crutchfield, R. S., "Experiments in the Use of Programed Instruction for the Facilitation of Creative Problem Solving," Programed Instruction, 4: 3-10, 1965.

3. Covington, M. V., Crutchfield, R. S., and Davies, L. B., The Productive Thinking Program, Series One: General Problem Solving, Brazelton Printing Company, Berkeley, California, 1966. Distributed by Educational Innovation, Box 9248, Berkeley, California 94719.

4. Crutchfield, R.S., "Creative Thinking in Children: Its Teaching and Testing," in O. Brim, R.S. Crutchfield, and W. Holtzman, Intelligence: Perspectives 1965, Harcourt, Brace, and World, Inc., New York, 1966, p. 33-64.

5. Getzels, J. W., and Jackson, P. W., Creativity and Intelligence: Explorations with Gifted Students, Wiley, New York, 1962.

6. Guilford, J. P. and Hoepfner, R., "Creative Potential as Related to Measures of IQ and Verbal Comprehension," Indian Journal of Psychology (1), 41: 7-16, 1966.

7. Klausmeier, H. J., "Effects of Accelerating Bright Older Elementary Pupils: A Follow Up," Journal of Educational Psychology, 54: 165-171, 1963.

8. Klausmeier, H. J., Harris, C. W., and Ethnathios, Z., "Relationships Between Divergent Thinking Abilities and Teacher Ratings of High School Students," Journal of Educational Psychology, 53: 72-75, 1962.

9. Klausmeier, H. J., and Teel, D., "A Research-based Program for Gifted Children," Education, 85: 131-136, 1964.

10. Klausmeier, H. J., and Wiersma, W., "Relationship of Sex, Grade Level, and Locale to Performance of High IQ Students on Divergent Thinking Tests," Journal of Educational Psychology, 55: 114-119, 1964.

11. Olton, R. M., and Crutchfield, R. S., "Developing the Skills of Productive Thinking," in P. Mussen, J. Langer, and M. Covington (eds.), New Directions in Developmental Psychology, Holt, Rinehart and Winston, New York, 1969, (in press).

12. Olton, R. M., Wardrop, J. L., Covington, M. V., Goodwin, W. L., Crutchfield, R. S., Klausmeier, H. J., and Ronda, T., "The Development of Productive Thinking Skills in Fifth-grade Children," Technical Report No. 34, Wisconsin Research and Development Center for Cognitive Learning, Madison, Wisconsin, 1967.

13. Ripple, R. E., and May, F. W., "Caution in Comparing Creativity and IQ," Psychological Reports, 10: 229-230, 1962.

14. Torrance, E. P., Guiding Creative Talent, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

15. Wallach, M. A., and Kogan, N., Modes of Thinking in Young Children: A Study of the Creativity-intelligence Distinction, Holt, Rinehart and Winston, New York, 1965.

ON TO THE CLASSROOM

A Guide for Preschool Teachers in Head Start-Type Programs of Compensatory Education

EDITED BY

Robert E. Clasen

200 pages $7.25 Hardcover and $5.75 Softcover

ON TO THE CLASSROOM deals with typical problems common to teachers of disadvantaged preschool children and contains unique suggestions for understanding and meeting the needs of these youngsters. The chapters are based on papers by well-qualified professors and professionals from the preschool education field which were originally presented to a group of Head Start teachers needing help in the various areas covered. The editor says, "Since these works were extremely useful to one group of teachers, they should be useful to others."

The book begins with a chapter which defines "culturally deprived" and offers a frame of reference for the thoughts and ideas presented in the remainder of the book. Each chapter was selected by Dr. Clasen on one criterion: *Does it contain information which our experience has shown that teachers need?* The chapters speak for themselves:

Creating a Learning Environment (numerous hints are given on how this learning climate can be created) (Chapter 2)

The Teacher, The Child and Head Start (the needs of children and a teacher's awareness are discussed) (Chapter 3)

Speech Language Acquisition and Language and Head Start (deal with language diagnosis and teaching strategies) (Chapters 4, 5)

From a Teacher's Point of View (a humorous and heart-rending day to day account of organizing, canvassing, and parent programming in Head Start, plus the happenings in a Head Start classroom from the first class day to the last—all taken from a teacher's log with her commentary and suggestions) (Chapter 6)

A Conversation with a Head Start (A.D.C.) Mother (reveals what the mother of a Head Start child experiences) (Chapter 7)

Programming for Parents (offers surprising views on what this is all about) (Chapter 8)

A Statement by Dr. Clasen summarizes the real purpose of ON TO THE CLASSROOM: "The fondest hope of each of us is that an idea shared through this medium may stimulate a change in a teacher's behavior for the benefit of a child."

PLEASE SEND ME THE INDICATED NUMBER OF COPIES OF: ______________________

ders

DEMBAR
EDUCATIONAL
RESEARCH
SERVICES, INC.

POST OFFICE BOX 1148 • MADISON, WISCONSIN 53701

NAME______________________
ADDRESS______________________
CITY______________________ STATE__________ ZIP__________

☐ I enclose a check for postpaid books ☐ Bill me and I'll pay the postage

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

FREE RECALL OF VERBAL MATERIAL IN THE CLASSROOM AS A FUNCTION OF METHOD OF PRESENTATION AND METHOD OF PRACTICE[1]

JAMES R. WHITMAN

Veterans Administration Hospital
Tacoma, Washington

ABSTRACT

Order in which stimuli were presented, and practice instructions were varied for 182 Ss receiving classroom practice on a free recall task with lists of nonsense words. The order within the list was a critical factor for the two levels of task difficulty used. Ss who studied the list gave more correct responses than did those who copied the list with the difficult material only. Ss then received additional practice with a second list. No facilitation in the free recall of the second list, attributable to the conditions which facilitated the learning of the first list, was found.

A CLASSIFICATION of the types of variables descriptive of the classroom presentation as a learning situation would include those which describe what is to be learned, how the task is presented, and the conditions under which the student participates in the presentation. Subject matter varies, of course, in difficulty; but an instructor is frequently making decisions regarding how the content of a lesson is to be presented, i. e., how the presentation is organized for the student. He may also have some control over the method used by the student in completing assignments.

In this study these factors were interpreted in terms of variables influencing free recall, with lists of nonsense words constituting the material to be learned. Nonsense words instead of textbook material were used since most students are unfamiliar with these; and what is more important, recall for this type of material involves the learning of the pattern of interrelationships (forms of internal structure, Garner 1962) existing between letter variables. Different methods for the presentation of stimuli also have been shown to influence the rate at which certain forms of internal structure can be learned (3). The task used accordingly is one involving the perception of relationships in the material learned, and the purpose of this study is to investigate the effect of performance of varying both the method for the presentation of such material and the method used by the learner. Ss received practice on a list of nonsense words, were tested for recall of these, and then received additional practice on a second list. The hypotheses tested were to the effect that (a) studying a list would result in a greater number of words recalled than would copying the words in the list; (b) grouping the words according to similarities in word structure would result in a greater number of words recalled than would an arrangement in which such groupings are avoided; and (c) the effects produced by studying and grouping the words would transfer to a second task for which studying and grouping procedures were not used. To add generality to the findings, the nature of the interrelationships between the letter variables, which constituted the learning task, was varied by using two levels of complexity.

METHOD

General Procedure

The experiment was conducted on a single day and during the regularly scheduled class times. The method of complete presentation was used. The lists of words to be learned and all instructions received by Ss were contained in a booklet form distributed to S. The first page of the booklet contained information on how the experiment was to be conducted. This information was read and explained to the class. Subsequent pages contained the material to be learned and all directions for practice (learning period) followed by a lined sheet with instructions for recording answers (answer period). All learning periods and test periods were timed and each lasted 90 seconds. A learning period followed by a test period constituted a trial. Two trials were given with the

S's first list, and these were followed immediately by one trial with the S's second list.

EXPERIMENTAL CONDITIONS

Eight experimental groups of twenty-two to twenty-three each were used. Each represented one of the possible combinations in a 2 x 2 x 2 experimental design. The experimental variables, each with two values, were: form of internal structure, i. e., level of complexity, order in which the words were listed, and instructions for practice. Form of internal structure was varied by using one list of four letter words in which all of the internal structure was in the form of a simple contingency between the letter values used in the first two positions of the words (List 1), and one list of four letter words in which all of the internal structure was in a form which did not include any simple contingencies between the letter values in the words (List 2).

Grouping was varied in the following manner. Words in a list were ordered either with the same first (and second insofar as possible) letters together (Grouped Order)or such ordering was avoided (Mixed Order). Both groupings were used for the first learning task. Only the Mixed Order was used for the second. The following instructions for practice were given for the different trials: For the Study Groups on Trial 1: ''Study the list below and try to find out the rule or system by which the words for the list were selected. The rule should be a statement which would help a person learn the list when he sees it. The order is not important. Do not try to remember the words--look for the rule or system. Do not write on this page.'' For the Study Groups on Trial 2: ''Below is the same list. This time study the list so that you can remember all of the words. Remember the rule. Do not write on this page.''

For the Practice Groups on Trials 1 and 2: ''Copy the words in the list below and try to remember them as you print them. Print the words in the order given, but do not try to learn the order.'' For all Groups on Trial 3: ''Below is a second and different list. Copy the list and try to remember all of the words--in any order.''

All experimental conditions were used with each class presentation. For each condition separate forms of a booklet were used. The classroom distribution of the forms was in a predetermined random order for blocks of eight booklets each and with the following restriction. A digit substitution task had been given to Ss at the beginning of the class session, and Ss received their booklets on the basis of a rank ordering of their scores on this task.

THE LISTS[2]

Lists 1 and 2 were used for the first learning task. List 1 was the following: BROZ, PLAJ, BRAZ, PLOJ, BROJ, PLAZ, BRAJ, and PLOZ. List 2 was: BROZ, PLAJ, BLAZ, PROJ, BLOJ, PRAZ, BRAJ, and PLOZ.

Lists 1a and 2a were used for the second learning task for the groups who received practice on List 1 and List 2 respectively. List 1a was: BLAZ, PROJ, BLAJ, PROZ, BLOZ, PRAJ, BLOJ, and PRAZ; and List 2a was: BRAZ, PLOJ, BLAJ, PROZ, BLOZ, PRAJ, BROJ, and PLAZ. The form of internal structure for List 1 is identical to that for List 1a; and that for List 2 is identical to that for List 2a.

SUBJECTS

Ss were 182 students in four undergraduate psychology courses. They had been told a week before the time of the experiment that an experiment in learning would be conducted during one of their class periods.

RESULTS AND DISCUSSION

S's score was the number of words correctly recalled in any order. On Trial 1, Ss who were instructed to study the list were required to state a rule. This rule was not scored. Those who copied the list were required to reproduce the list.[3] Comparisons between groups were on the basis of number of S's correctly recalling all eight words in the list on Test Trial 2 which was the first test trial in which recall was required for all groups. These data are shown in Table 1. Also the data, expressed as percentages of Ss, and grouped according to the experimental variables, are shown in Figure 1. Chi-square, corrected for continuity, was used to determine the significance of differences; and the $\underline{p} < .05$ level was used.

The two lists differed in level of difficulty, and comparisons were made for each list separately. Also comparisons were made between pairs of groups as shown in Figure 1. For the less difficult list (List 1), the number of Ss giving all correct responses was greater with the Grouped than with the Mixed presentation, but the difference on this measure between Study and Copy Method was not significant. For List 2, the more difficult list, the difference between the Grouped and Mixed presentation was also significant and in the same direction. With this learning task, however, more Ss gave all correct responses with the Study Method than with the Copy Method.

Ss received practice on a second list. S's scores after practice on this list are shown in Table 2. For List 1a, comparisons were made on the same basis as that used for Lists 1 and 2 above; but for List 2a the comparisons were on the basis of mean number of correct responses per group since only twelve out of ninety Ss reached the criterion of all correct responses; and an analysis of variance and the $p < .05$ level were used with these data. Also, comparisons were made for each list separately. The differences between the groups receiving practice on List 1a were not significant; also, the differences between the groups receiving practice on List 2a were not significant. The finding, therefore, is that the effects produced by studying the grouping of the words in the first list were not transferred to a second task for which studying and grouping procedures were not used.

These results lead to two conclusions. The first is that for a task requiring the learning of interrelationships between variables the order or arrangement in which the material is presented is a critical factor. This result extends the findings of Whitman and Garner (3) who found that different orders for presentation of positive and negative instances can increase the difficulty of a concept learning task. The study method, i. e., looking for a rule rather than copying the material while trying to memorize, produced more correct responses with the difficult list only. The second conclusion, therefore, is that the efficiency of the study method is related to the level of difficulty of the learning task.

The transfer effect reflected in the performance

TABLE 1

PERCENT OF Ss GIVING ALL CORRECT RESPONSES (ROMAN) AND MEAN PERCENT CORRECT (ITALIC) AFTER PRACTICE ON THE FIRST LEARNING TASK. GROUPING IS BY MATERIAL IN LEARNING TASK, ARRANGEMENT OF THE MATERIAL, AND BY PRACTICE INSTRUCTIONS[a]

		Material in Learning Task			
		List 1		List 2	
Practice Instructions:		Study	Copy	Study	Copy
Arrangement of Material	Grouped	87	96	61	23
		97	*99*	*72*	*68*
	Mixed	61	61	26	14
		88	*92*	*61*	*52*

[a]N = 23 for all groups except for the two groups copying List 2, where N = 22.

FIGURE 1

PERCENTAGE OF Ss GIVING ALL CORRECT RESPONSES FOR COMPARISONS BASED ON THE EXPERIMENTAL VARIABLES (N = 92 FOR LIST 1; N = 90 FOR LIST 2.)

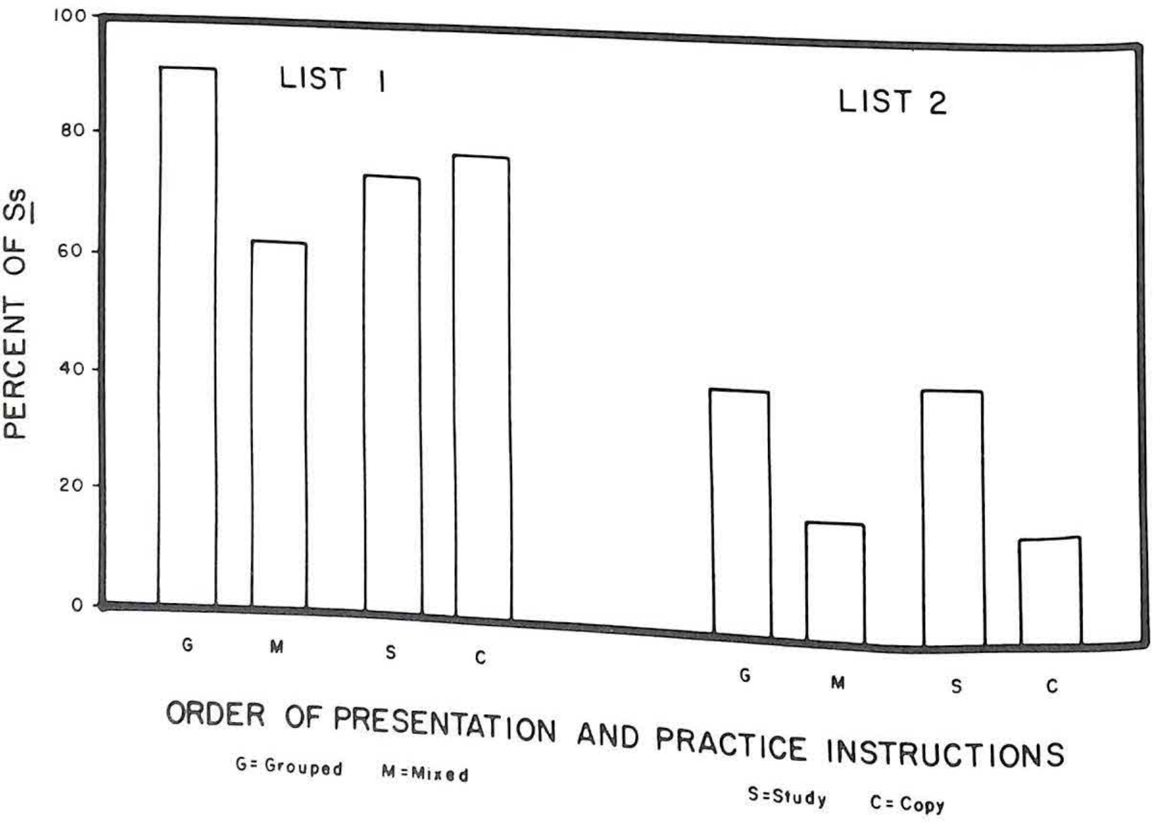


TABLE 2

PERFORMANCE OF GROUPS ON TRANSFER TASK PERCENT OF Ss GIVING ALL CORRECT RESPONSES (ROMAN) AND MEAN PERCENT CORRECT (ITALIC) AFTER LEARNING TRIAL. GROUPING IS BY EXPERIMENTAL CONDITIONS UNDER WHICH *FIRST* LIST WAS LEARNED [a]

		Material in Learning Task			
		List 1		List 2	
Practice Instructions:		Study	Copy	Study	Copy
Arrangement of Material	Grouped	70	70	17	5
		84	*79*	*57*	*56*
	Mixed	70	61	13	18
		83	*82*	*59*	*56*

[a] N = 23 for all groups except for the two copying List 2 as the first task, where N = 22.

of the different groups on the second learning task is more difficult to interpret. It is of considerable importance to determine the extent to which the results obtained are a function of the amount of practice, i. e., number of lists learned.

FOOTNOTES

1. A portion of this paper was presented at the 75th Annual Convention, APA, 1967.
2. For a detailed description of the Lists used in this study see Garner and Whitman, 1965.
3. A significantly greater number of Ss copying the ordered List 1, as contrasted to the mixed List 1, gave all correct responses on this trial ($X^2 = 7.393$, df = 1). No S copying List 2 gave all correct responses.

REFERENCES

1. Garner, W. R., *Uncertainty and Structure as Psychological Concepts*, Wiley, New York, 1962.
2. Garner, W. R., and Whitman, J. R., "Form and Amount of Internal Structure as Factors in Free-recall Learning of Nonsense Words," *Journal of Verbal Learning and Verbal Behavior*, 4:257-266, 1965.
3. Whitman, J. R., and Garner, W. R., " Concept Learning as a Function of Form of Internal Structure," *Journal of Verbal Learning and Verbal Behavior*, 2:195-202, 1963.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

IS THE PHONETIC GENERALIZATION HYPOTHESIS IN SPELLING VALID?[1]

ALBERT H. YEE

University of Wisconsin, Madison

ABSTRACT

This study was conducted to help settle a major controversy in spelling instruction concerning the phonetic generalization hypothesis. It utilized Solomon's Four-Group Experimental Design for 2 x 2 x 2 factorial analyses--girls versus boys, pretesting versus nonpretesting, and phonetic instruction versus no phonetic treatment. Dependent variables were two spelling test scores of pupils in 106 Grades 2-6 classes: (1) spellings consistent with two phonetic generalizations taught to treatment groups and (2) spellings unrelated to the two phonetic generalizations. In order of most significant effects found, results showed girls superior to boys, pretesting better than nonpretesting, and questionable advantage for phonetic instruction over no phonetic instruction. Test-Study methods appear to be more promising than phonetic approaches. Further research is necessary to ascertain the definite value of Test-Study methods.

A CONTROVERSY between leading authorities in English-American orthography has continued for many years concerning the question: Can spelling proficiency be improved through greater use of spelling generalizations? According to one view, spelling instruction should be based primarily on drill and practice with the use of spelling generalizations regarded only as one supplemental aid. In contrast, some authorities feel that English-American orthography has a sufficient degree of sound-to-letter relationship to justify using phonetic approaches as the primary instructional method in spelling programs. Yee (23) has written a critical review of this controversy.

The controversy boils down to the question: Which instructional method should predominate over others? According to Horn (11:432), "When the evidence, on both the consistency and the irregularities of English spelling, is critically and realistically assessed, little justification is found for the claim that pupils can arrive deductively at the spelling of most words they can pronounce." Horn believed that some emphasis on phonetic approaches may be justified, but only as a secondary aid, not as a substitute for the systematic study of individual words in spelling lists (10:24). In contrast, Hanna and others (7) have suggested that a spelling program based upon methods emphasizing oral-aural cues might well prove to be superior over those methods which use primarily visual and drill approaches. Hodges (9:634) has argued: "Sequential training in helping children to listen for phonemes in relation to their placement in words, to the way in which they are enunciated, and to the basic regularity with which they are represented by graphemes in writing creates a basis for effective spelling power."

Although spelling has become a content area of less concern than others, such as reading, mathematics, and social studies, it continues to be an important curricular area. A factor analytic study by Allen and Ager (1) found that spelling is an independent skill and "should receive specific attention and instruction in the school curriculum" (1:159). The general public perceives the ability to spell as a mark of educational achievement and often evaluates the success of schools by the spelling proficiency of graduates. Spelling is important because it involves the purchase of millions of expendable workbooks each school year and has become big business. As expected, such purchased materials reflect varying positions between the two points of view. Also, despite its lower status as a content area, a significant part of daily school work continues to be spent in spelling. Therefore, the controversy directly affects all elementary school teachers and their pupils.

We conducted this study to ascertain the value of providing phonetic generalizations in spelling instruction. Thus, we hypothesized that spelling scores of pupils taught phonetic generalizations as prescribed by a leading spelling series will be greater than scores of pupils not taught phonetic generalizations.

In establishing the hypothesis, we attempted to follow Platt's (18) advice that the best scientific procedure in research is to build inferential thinking to the point where different courses of action (i. e., as in inferential trees with branching alternatives) need to be tested by crucial experiments. In the area of spelling instruction, it is important to know if phonetic generalizations do or do not make a significant difference in spelling achievement, because the spelling controversy has in fact reached the branching point. Empirical research needs to ascertain yea or nay to the generalization hypothesis. Thus, it is important to see that differing implications follow either outcome.

METHOD[2]

Subjects

In the Fall Semester of 1965, all 106 elementary student teachers at the University of Texas and their pupils in grades 2 to 6 in Austin, Texas schools were recruited for this study. Grade one classes were omitted from the study because they lacked extensive formal spelling instruction. Since student teacher assignments in elementary grades 2-6 were in all types of schools in the district, the distribution of elementary grade classes included a fairly random representation of pupils from the school district. The twelve schools in which student teachers were located had at the time of testing a total average daily attendance of 6,063 pupils of which 2,839 (posttest N) pupils participated in the study.

It was possible to evaluate the sample drawn for this study with objective information from University and school records in relation to student teachers' and pupils' characteristics. From a survey of such information, the sample was judged representative of typical teacher candidates and pupils in urban centers and offered no systematic bias. This sample, therefore, makes it possible to generalize results and conclusions more readily to similar situations elsewhere.

Factorial Design

The student teacher-class pairs were randomly grouped into the experimental framework of Solomon's Four-Group design (21), which Campbell and Stanley (3:194-195) have recommended as a "true experimental design."

Subjects were randomly assigned to one of four groups. Group A represented the experimental group where pupils were given pre- and posttests and instruction in two selected phonetic generalizations in the interval between testing occasions. Group B represented the first control group where pupils were given pre- and posttesting but no phonetic instruction. Groups C and D received only posttesting, but treatment group C was taught phonetic rules and their application prior to testing and control group D was not.

Therefore, the study's experimental design is as follows:

		Pretest		Posttest
R →	Experimental Group A	0_1	x	0_2
	Control Group B	0_3		0_4
	Control Group C		x	0_5
	Control Group D			0_6

Where R = randomization, 0 = test observations, and x = treatment, i.e., instruction in phonetic generalizations.

The posttest scores of pupils in the four groups were converted to class means, and analysis of variance was used to test the statistical null hypothesis: $0_2 = 0_4 = 0_5 = 0_6$. With two levels each for the independent variables of pupil's sex (X), pretesting (Y), and phonetic treatment (Z), a 2 x 2 x 2 factorial analysis was conducted for each grade level. Simultaneous analysis of all grade levels together was not done, because the tests were tailored for each grade level and cannot be considered parallel.

By designing our experiment with independent variables other than Z, we reasoned that the competing potencies of sex differences (X) in verbal ability and of pretesting (Y) need to be considered. Sex differences have been a familiar source of significant variance in verbal ability. For the latter, there has been strong evidence to support the Test-Study and Corrected-Test approaches (e.g., 5, 12), which place great weight on pretests to help pupils find words that require their concern and study. Such methods, however, have been slow to gain prominence in spelling and class instruction. Perhaps the lag comes from the common conformity to spelling workbooks.

Dependent Variables and Treatment

A spelling test was developed for each grade level with three categories of words: (1) Phonetic spellings that followed two phonetic generalizations taught the particular grade level; (2) Tricky spellings that nearly followed the generalizations taught, but they would be incorrectly spelled if the rules were followed literally; and (3) No relationship spellings that lacked relationship to the generalizations taught treatment groups. Thus, each category of test words formed a separate dependent variable. Second thoughts made us believe that a clear distinction between tricky and no relationship words was not fully satisfactory for all grade levels. The F results were essentially identical for the categories, and no real gain was achieved by separate analyses. Reliability would be increased by combining the two word lists, thus they were classed together to form one word category which we termed unrelated words, i.e., words that cannot be successfully spelled by applying the generalizations.

Table 1 presents the test words for each grade level by showing the spellings consistent with the phonetic generalizations taught in lower-case letters and the spellings unrelated to the generalizations in upper-case letters.

A noted authority in spelling, T. D. Horn, helped prepare and check the test words and placement in the categories. Table 1 does not list the words as they were presented in the test. The words were randomly distributed in tests and are grouped according to their categories in Table 1 for the convenience of the reader.

Lessons for two representative generalizations were chosen for each grade level from Kottmeyer and Ware's (13) Basic Spelling Goals series. To maximize treatment effects, daily lessons were taken exactly from the series with the standard week's work given to each generalization. Words used in the lessons were not used in the above spelling tests. The lessons of the popular Kottmeyer and Ware series exemplify the phonetic approach. Through regular use of the series, pupils of the school district would be sensitive to phonological regularities and instruction emphasizing them. Thus, classroom conditions

TABLE 1

GRADE LEVEL TEST WORDS

2	3	4	5	6
three	babies	lately	music	average
eat	factories	totally	pilot	garage
feet	coming	electricity	glider	notion
ride	cherries	gravity	title	revolution
rode	having	colony	syllable	rotation
make	making	freely	article	projection
tail	daisies	queenly	radar	dangerous
seat	countries	lively	locate	engage
ate	riding	exactly	cable	fudge
goat	caring	century	startle	edge
tea	families	discovery	giggle	promotion
deep	ponies	equally	lemon	fiction
game	funnies	uselessly	lazy	grudge
bike	bodies	evenly	above	damage
nine	closing	geography	against	invention
gave	leaving	army	pebble	sponge
seed	moving	fifty	bugle	dodge
need	skating	poetry	scribble	attraction
kite	skies	purely	ahead	conjunction
stove	jellies	library	grumble	courage
cake	ladies	fairly	rectangle	objection
soap	using	unkindly	triangle	selection
train	writing	usefully	sugar	recreation
cone	hiding	loosely	potato	college
take	candies	loudly	stable	ADJECTIVE
neat	bunnies	hobby	triple	INJURE
near	shaking	proudly	reptile	OBJECT
pail	taming	energy	banana	FORTUNE
came	berries	MOVIE	PRISON	SOLDIER
joke	storing	PRAIRIE	AVENUE	MAJOR
HEAD	PLAYS	RECIPE	EXERCISE	ANGELIC
THREE	PAYS	ABLY	OVEN	LONGITUDE
VIEW	FREEING	MACARONI	THEATER	SERGEANT
COME	LYING	LUCKILY	TOTAL	COMPLEXION
LOVE	TURKEYS	UNSTEADILY	CRUEL	PROFESSION
ROOT	DYING	TRULY	FINAL	MISSION
SAID	BUYS	DISAGREE	SHOVEL	EXCLUSION
HAVE	ENJOYS	WHOLLY	SPECIAL	OCCASION
ONE	HOEING	VALLEY	PARCEL	INCLUSION
PIECE	SHOEING	HOLIDAY	PUPIL	FISSION

were favorable for the treatment effects of this study. The generalizations chosen for each grade level are as follows:

Grade 2: (1) When two vowels are together in short words, the first vowel "says its name" and the second is silent (Lesson 14 and 15). (2) When I see a vowel, then another letter (consonant), and then an e at the end of a short word, I hear the name of the vowel and the e is silent (Lesson 16).

Grade 3: (1) For words which end in silent e, we drop the e before we add -ing (Lesson 20). (2) Many words which end in y form the plural by changing the y to i and adding es (Lesson 26).

Grade 4: (1) The sound of i at the end of a word is usually spelled by the letter y; when y follows a vowel, as in words like key, the y is silent (Lesson 27). (2) -ly is added to root words without changing the spelling of that root with the exception of truly (Lesson 31).

Grade 5: (1) Many words end with the -el sound. Words ending with this sound often have the -le spelling (Lesson 26). (2) Words in the vowel-consonant-vowel pattern are usually divided between the first vowel and consonant, the second vowel going with the consonant (Lesson 32).

Grade 6: (1) The soft g or j sound at the end of a word is usually spelled ge or dge as in rage and judge (Lesson 13). (2) The noun suffix -tion is usually pronounced shen, -sion as zhen, and -ion as yen (Lesson 21).

Procedures in Testing and Treatment

Mechanical aids, such as tape recorders, were not used to help control treatment and test conditions. Providing the more natural classroom situation was deemed less of a threat to internal validity than the possibility of ''novelty effect'' and limitations in pupil feedback during treatment sessions. To help ensure uniformity in test administration, each spelling test was printed according to the normal procedures for spelling tests. In other words, the tests presented the word to be pronounced, a sentence in which the word was used, and then required that the word be read again.

We took care to avoid contamination of the data by the student teachers by informing the groups of student teachers only what was necessary for them to know at each stage. Three special orientation sessions were held for the following purposes and groups: (1) pretesting of Groups A and B; (2) treatment for Groups A and C; and (3) posttesting of all four groups. For example, a few days after the pretests of Groups A and B were collected, the teachers of Groups A and C met together in a group meeting where they were given daily lesson plans for 2 weeks' classwork. They were instructed to teach the lessons as printed and demonstrated to them in the group meeting. After a period of about 4 weeks after pretesting, posttests were administered to all groups. The use of class means in analyses also helped control the teacher variable.

Thus, an interval of about 2 weeks separated the completion of phonetic instruction in treatment groups and posttesting. This time interval was provided to allow an objective assessment of the treatment effect more independent of the treatment sessions. The treatment effect would be maximal immediately following phonetic instruction, and we did not want to invalid the posttests by the influence of immediate recall and mental set. Since the generalization hypothesis assumes long-term effect, the 2-week interval seems to be more than fair in terms of forgetting and reasonable in terms of overcoming immediate treatment effect.

RESULTS

Pretest Scores

To ascertain if there was any bias resulting from the randomization process, the difference between pretest scores for Groups A and B for each grade level was tested with 2 x 2 analysis of variance. The factors were pupil sex and group (Group A vs. Group B). The only significant results obtained were in Grade 6 where girls scored higher than boys with phonetic and unrelated words. Since no difference between Groups A and B was significant (most of the F ratios were less than one), the experimental group and the first control group was equivalent in respect to their responses to the dependent variables prior to treatment sessions.

Reliability

The internal consistency of the pupils' scores was estimated with Guttman's (6) L_4 formula with odd-even split-halves. All coefficients for whole class, grade-level total test and phonetic word scores

TABLE 2

MEANS AND STANDARD DEVIATIONS FOR PHONETIC WORDS

Grade Level	Group	Class N	Girls		Boys	
			M	SD	M	SD
2	A	5	20.38	1.74	16.78	3.17
	B	5	21.97	5.44	18.66	5.63
	C	6	20.06	1.74	17.67	4.50
	D	6	19.55	2.16	16.65	1.58
3	A	6	19.07	2.39	13.98	3.43
	B	5	16.60	3.42	13.50	3.26
	C	4	17.31	1.85	13.99	3.36
	D	5	11.90	3.48	9.66	3.96
4	A	5	15.36	3.14	14.79	1.40
	B	6	14.90	3.17	14.73	2.17
	C	4	13.38	3.00	14.07	3.07
	D	5	13.78	2.49	10.95	2.70
5	A	6	18.01	5.31	16.61	5.58
	B	5	19.52	6.47	20.10	1.46
	C	7	20.39	2.01	18.10	1.14
	D	5	18.98	7.08	16.57	4.37
6	A	5	18.17	1.09	15.06	2.42
	B	5	19.19	2.18	17.21	1.72
	C	5	18.81	1.32	15.73	2.10
	D	6	16.65	3.53	14.58	3.61

TABLE 3

ANOVA RESULTS FOR PHONETIC WORDS

Grade	Source	df	MS	F
2	Pupil sex (X)	1	110.71	8.86**
	Pretesting (Y)	1	7.47	.60
	Treatment (Z)	1	1.30	.10
	XY	1	.77	.06
	XZ	1	.08	.01
	YZ	1	21.05	1.69
	XYZ	1	.04	.004
	Within Groups	36	12.49	
	Between Groups	7	20.20	
	Total	43	13.75	
3	X	1	115.75	11.17**
	Y	1	64.81	6.25**
	Z	1	98.38	9.49**
	XY	1	4.22	.41
	XZ	1	5.77	.56
	YZ	1	28.24	2.72
	XYZ	1	.49	.05
	Within Groups	32	10.37	
	Between Groups	7	45.38	
	Total	39	16.65	
4	X	1	5.03	.70
	Y	1	35.35	4.91
	Z	1	6.38	.89
	XY	1	1.12	.17
	XZ	1	5.97	.83
	YZ	1	2.96	.41
	XYZ	1	9.38	1.30
	Within Groups	32	7.20	
	Between Groups	7	9.47	
	Total	39	7.61	
5	X	1	21.59	1.33
	Y	1	.02	.002
	Z	1	2.98	.18
	XY	1	10.61	.66
	XZ	1	2.45	.15
	YZ	1	44.54	2.75
	XYZ	1	3.07	.19
	Within Groups	38	16.19	
	Between Groups	7	12.18	
	Total	45	15.57	
6	X	1	68.54	11.04**
	Y	1	9.74	1.57
	Z	1	.01	.002
	XY	1	.001	.0002
	XZ	1	2.97	.48
	YZ	1	27.39	4.41
	XYZ	1	.001	.002
	Within Groups	34	6.21	
	Between Groups	7	15.52	
	Total	41	7.80	

*$p < .025$ **$p < .01$

were in the 90's. With fewer words than the phonetic word lists, the unrelated word scores had lower but adequate rs averaging about .75 with the lowest r being .69 for Grade 4. Separate results for girls and boys closely resembled those for whole classes. Such results indicate the satisfactory reliability of pupils' measures.

Phonetic Test Words

Table 2 presents posttest phonetic word means and standard deviations for girls and boys in the four groups. Some heterogeneity of variance can be observed in some but not all grade level SDs, but this

TABLE 4

MEANS AND STANDARD DEVIATIONS FOR UNRELATED WORDS

Grade Level	Group	Class N	Girls		Boys	
			M	SD	M	SD
2	A	5	6.17	.31	5.03	.83
	B	5	6.61	1.28	5.90	1.32
	C	6	6.45	.70	5.53	.82
	D	6	5.98	.57	4.95	.35
3	A	6	4.92	.78	3.53	.61
	B	5	5.07	1.62	3.70	.99
	C	4	4.41	.50	3.47	.68
	D	5	4.51	.67	2.99	.60
4	A	5	3.68	.95	3.27	.61
	B	6	3.78	1.13	3.49	.55
	C	4	3.21	1.11	3.57	.76
	D	5	3.66	.89	2.41	.71
5	A	6	7.06	2.40	6.75	2.20
	B	5	7.52	.74	7.65	1.14
	C	7	8.01	.84	7.32	.92
	D	5	7.95	2.44	6.43	1.66
6	A	5	9.34	1.06	6.99	1.86
	B	5	10.15	1.77	8.40	.91
	C	5	9.26	.70	7.45	1.67
	D	6	7.71	1.75	6.46	1.95

problem is offset by establishing a higher level overall required for statistical significance, i.e., .025 instead of .05 (14:86).

The effect of heterogeneity of variance upon results was tested in preliminary analyses by equalizing all Ns by randomly selecting out classes in groups larger than the cell with the lowest class N (8:379). Since those F results were equivalent to those obtained with unequal Ns, we report the latter results computed with Veldman's (22) AVAR-23 program which can handle triple-classification analysis of variance (ANOVA).

As can be seen in Table 2, all but two of the differences between means for the pupil sex factor favor the girls. In Table 3, results for the factor of pupil sex show three F ratios surpassing the .01 level for differences great enough to be statistically significant in Grades 2, 3, and 6.

For the pretesting factor (Y), one significant F ratio was obtained (Grade 3), and all differences between means show pretested groups superior to nonpretested groups.

Only one significant F ratio (Grade 3) was also found for the treatment factor (Z); but unlike results for Y, the differences between means indicate no uniform pattern suggesting a direction of effect.

Thus, in accounting for variance in pupils' phonetic spellings, the results indicate that the most potent contributor to variance is pupil's sex, with the milder positive effect of pretesting second, and the uncertain effect of phonetic instruction least significant. No significant two- or three-way interaction effects were found with phonetic words.

Unrelated Test Words

Table 4 presents posttest unrelated word means and standard deviations for girls and boys in the four groups. All but two of the means for girls are greater than those for boys.

Table 5 ANOVA results with unrelated words also show that factor X accounts for more of the variance in spelling behavior than Y or Z. These highly significant F ratios indicate the superiority of spelling ability by girls over boys.

No significant F ratio was found for the effects of Y and Z. While the pretested groups scored higher than nonpretested groups in all grade levels, there was no advantage in contrasts between phonetic treatment and non-treatment groups. Such distinctions are evidenced by all F ratios for the Z factor being less than one and three of the five F ratios for the Y factor being greater than one.

The significant YZ interaction effect for sixth graders' responses to unrelated words (F of 6.14, $p < .017$) reflects higher spelling scores achieved with either Y or Z than with both main effects present or absent. A similar YZ interaction effect reaches the .04 level with phonetic words (Table 3) in grade 6. Although it is considered statistically nonsignificant, the ranking of the means were identical to those found with unrelated words. Pretesting alone produced the highest means, so the pattern of

TABLE 5

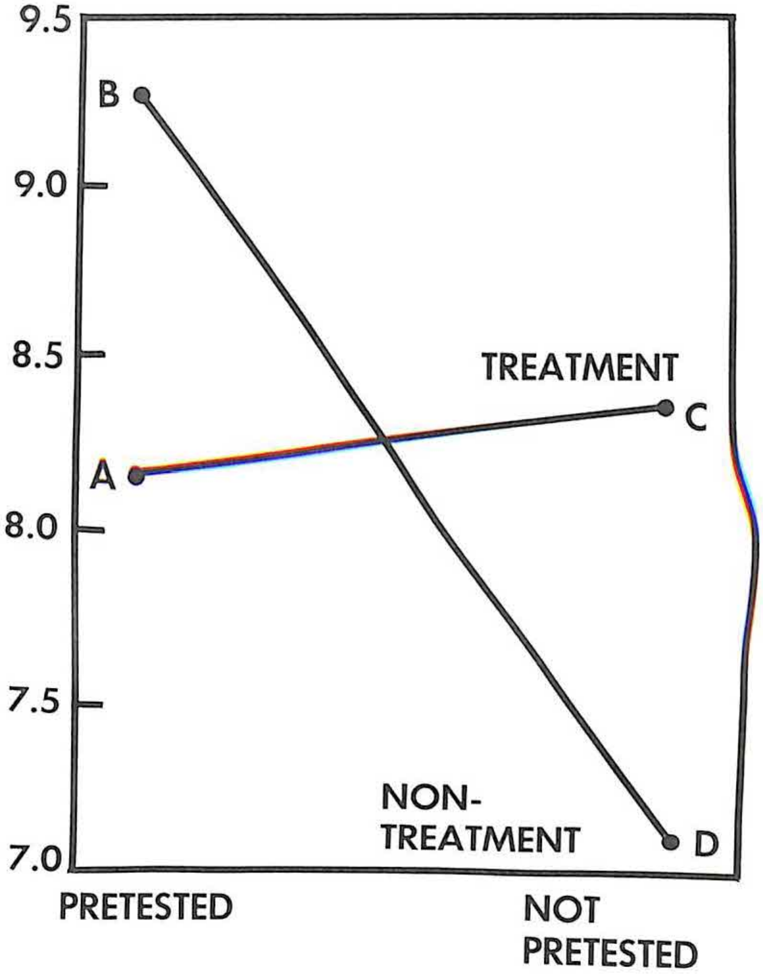
ANOVA RESULTS FOR UNRELATED WORDS

Grade	Source	df	MS	F
2	Pupil sex (X)	1	9.64	13.62**
	Pretesting (Y)	1	.39	.56
	Treatment (Z)	1	.07	.10
	XY	1	.002	.003
	XZ	1	.09	.12
	YZ	1	3.64	5.14
	XYZ	1	.17	.24
	Within Groups	36	.71	
	Between Groups	7	2.00	
	Total	43	.92	
3	X	1	16.67	21.67**
	Y	1	2.09	2.72
	Z	1	.002	.003
	XY	1	.06	.07
	XZ	1	.18	.24
	YZ	1	.31	.40
	XYZ	1	.23	.30
	Within Groups	32	.77	
	Between Groups	7	2.79	
	Total	39	1.13	
4	X	1	1.50	2.03
	Y	1	1.09	1.48
	Z	1	.11	.14
	XY	1	.03	.04
	XZ	1	1.30	1.76
	YZ	1	.59	.80
	XYZ	1	1.90	2.57
	Within Groups	32	.74	
	Between Groups	7	.93	
	Total	39	.77	
5	X	1	3.99	1.45
	Y	1	.38	.14
	Z	1	.12	.04
	XY	1	2.87	1.04
	XZ	1	.11	.04
	YZ	1	3.75	1.36
	XYZ	1	1.11	.40
	Within Groups	38	2.76	
	Between Groups	7	1.76	
	Total	45	2.60	
6	X	1	33.42	13.92**
	Y	1	10.41	4.34
	Z	1	.07	.03
	XY	1	.71	.30
	XZ	1	.88	.37
	YZ	1	14.74	6.14*
	XYZ	1	.001	.0003
	Within Groups	34	2.40	
	Between Groups	7	8.61	
	Total	41	3.46	

*$p < .025$, **$p < .01$

FIGURE 1.

SIGNIFICANT YZ INTERACTION IN SIXTH GRADERS' SPELLINGS OF UNRELATED WORDS.



group means goes as follows: B> C > A > D. Although results for other grades lack significant Fs for YZ interaction, the same hierarchial order can be generally found to support the predominance of such a pattern. Figure 1 illustrates this predominant pattern with sixth graders' XY results with unrelated words.

Figure 1 illustrates the significant departure of curves from parallelism. Group B, which received pretests but no phonetic treatment, scored highest; Group C scored second highest and was the treatment group which was not pretested; third in line is Group A, the experimental group which received both pretesting and treatment; and last comes Group D which received only posttesting; i. e., B > C > A > D. This resulting pattern indicates that either main effects, Y or Z, benefited pupils' spelling scores more than both of them together; but Y and Z together produced better results than when both were absent.

DISCUSSION AND CONCLUSIONS

The hypothesis that phonetic instruction provides a positive difference in spelling achievement was not confirmed in this study. The Solomon Four-Group design this study utilized makes it possible to suggest that pretesting may be more beneficial to spelling proficiency than phonetic instruction. If we had not investigated the effect of pretesting, the results would tend to support the generalization hypothesis. This can be seen in Tables 2 and 4 where most of the means for Group C are greater than those of Group D.

The predominant pattern of group means goes in the following general order: B > C > A > D. Since either pretesting in Group B or phonetic instruction in Group C appears to be generally superior to both effects together (Group A), a choice between pretesting and phonetic instruction in favor of the former seems feasible and deserves further study. The potential benefits of pretesting appear to go beyond the range of phonetically-regular words. Thus, the results support and complement findings for the Test-Study methods which emphasize pretesting procedures.

Let us attempt to explain such findings: Pretesting may create a mental set toward words that is sufficient to produce gains over time. Perhaps the mental set produces a dissonance process (4) which requires tension-reduction by the pupil's learning words that were recognized as difficult. Phonetic instruction alone may provide the pupil some preparatory process to attack unfamiliar words and thus help him perform slightly better than those with or without Y and Z together. But used as blanket rules, they lend themselves to more errorful spelling by perhaps creating less dissonance or concern on the part of pupils to check and concentrate upon the spelling of individual words.

In other words, there is a subtle difference between successful application of generalized rules for a number of different words and specific focus upon the correct and peculiar spelling of each word. Considering the irregularities of American-English spellings, we think the latter approach provides greater probability for success. For Group A, there may be a confounding of influences, whereby Y and Z cancel out each other. Further research is necessary to ascertain the validity of such findings. If Y and Z effects are found to conflict in later research, it would be important to learn why they do.

The most potent source of variance proved to be pupils' sex. It should not be surprising that girls spell better than boys do, since Rice (20:169) noted this difference long ago. This finding is also consistent with sex differences typically found with standardized achievement and intelligence scales. According to Anastasi (2:493):

> In general, girls surpass boys in those school subjects depending largely upon verbal abilities, memory, and perceptual speed and accuracy. Boys excel in those subjects that call into play numerical reasoning and spatial aptitudes, as well as in certain "information" subjects such as history, geography, and science.

We need to accept sex differences in spelling achievement for what they really are. More research is beginning to show that teachers often do not adequately fulfill the cognitive needs and abilities of boys and tend to punish them for resisting feminine-oriented behavior and interests (15, 19). It may not be worthwhile to provide differential spelling programs according to pupil sex, because many other factors contribute to individual differences at the classroom level. But, for the benefit of boys as well as girls, it would be helpful if teachers and language arts specialists who author spelling handbooks implemented more individualized procedures into their spelling programs that are based on intrinsic motivation and meaningful purposes.

Instead of either-or planning of instruction, it seems more realistic to determine priority among methods in spelling programs after specific behavioral objectives have been stated and carefully evaluated in terms of what comprises efficient spelling behavior. Then alternative approaches to handle the differing spelling needs of pupils can be tested. With such preliminaries, systematic instruction can be developed to better effect such behavior (16, 17).

Since the positive effect of ten lessons could not be detected in a pattern of superior spelling by treatment groups or in any significant Fs, the predominant use of generalizations for spelling instruction appears to be less useful than Test-Study methods. Test-Study methods appear to provide more efficiency and economy than phonetic instruction.

FOOTNOTES

1. The Research Committee of the Graduate School, University of Wisconsin, Madison provided research support for most of this work. The National Science Foundation also helped through their grant to the University of Wisconsin Computing Center. The Research and Development Center for Teacher Education, University of Texas, Austin supported data processing assistance for preliminary analyses. The author thanks Carl Personke for a critical reading of an earlier draft, but absolves him of responsibility for any shortcomings of the present version.

2. Loyal and helpful assistants made this study possible: Ethel Wiest led the group meetings and assumed major responsibility for data collection. H. Albert Napier, William Geeslin, Thomas Kriewall, and Troy Klein provided data processing and programming services with assistance from Janice Willenborg and Sue Sisk.

REFERENCES

1. Allen, Doris, and Ager, J., "A Factor Analytic Study of the Ability to Spell," Educational and Psychological Measurement, 25:153-161, 1965.

2. Anastasi, Anne, Differential Psychology: Individual and Group Differences in Behavior, (3rd ed.), Macmillan, New York, 1960.

3. Campbell, D. T., and Stanley, J. C., "Experimental and Quasi-experimental Design for Research on Teaching," in N. L. Gage (Editor), Handbook of Research on Teaching, Rand McNally, Chicago, 1963, pp. 171-246.

4. Festinger, L., A Theory of Cognitive Dissonance, Stanford University Press, Stanford, California, 1957.

5. Gates, A. I., "An Experimental Comparison of the Study-test and Test-study Methods in Spelling," Journal of Educational Psychology, 22: 1-19, 1931.

6. Guttman, L., "A Basis for Analyzing Test-retest Reliability," Psychometrika, 10: 255-282, 1945.

7. Hanna, P. R., Hanna, Jean S., Hodges, R. E., and Rudorf, E. H., Jr., Phoneme Grapheme Correspondences as Cues to Spelling Improvement, U. S. Office of Education Cooperative Research Program Monograph 16, Government Printing Office, Washington, D. C., 1966.

8. Hays, W. L., Statistics for Psychologists, Holt, Rinehart, and Winston, New York, 1965.

9. Hodges, R. E., "The Psychological Bases of Spelling," Elementary English, 42:629-635, 1965.

10. Horn, E., Teaching Spelling, American Educational Research Association, Department of Classroom Teachers Research Pamphlet, National Education Association, Washington, D. C., 1963.

11. Horn, E., "Phonetics and Spelling," Elementary School Journal, 57:424-432, 1957.

12. Horn, T. D., "The Effect of the Corrected Test on Learning to Spell," Elementary School Journal, 47:277-285, 1947.

13. Kottmeyer, W., and Ware, Kay, Spelling Goals, Webster Publishing Company, St. Louis, 1960.

14. Lindquist, E. F., Design and Analysis of Experiments in Psychology and Education, Houghton-Mifflin, Boston, 1953.

15. McNeil, J. D., "Programed Instruction Versus Usual Classroom Procedures in Teaching Boys to Read," American Educational Research Journal, 1:113-119, 1964.

16. Personke, C., and Yee, A. H., "A Model for the Analysis of Spelling Behavior," Elementary English, 43:278-284, 1966.

17. Personke, C., and Yee, A. H., "The Situational Choice and the Spelling Program," Elementary English, 45:32-37, 40, 1968.

18. Platt, J. R., "Strong Inference," Science, 146: 347-353, 1964.

19. Pollack, J. H., "Are Teachers Fair to Boys?" Today's Health, 46:20-25, 1968.

20. Rice, J. M., "The Futility of the Spelling Grind," The Forum, 23:163-172, 1897.

21. Solomon, R., "An Extension of Control Group Design," Psychological Bulletin, 46:137-150, 1949.

22. Veldman, D. J., Fortran Programming for the Behavioral Sciences, Holt, Rinehart and Winston, New York, 1967.

23. Yee, A. H., "The Generalization Controversy in Spelling Instruction," Elementary English, 43: 154-161, 1966.

THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 37, Number 4, Summer 1969)

THE USE OF MARKOV CHAINS TO ESTIMATE EFFECTS OF FEEDBACK ON DECISION-MAKING

GORDON WELTY
Chatham College, Pittsburgh

ABSTRACT

This research paper utilizes Billingsley's work on Markov processes in the estimation of behavior of the rational decisionmaker.

IN an earlier note (1), we introduced a reaction function R to describe the response of a rational manager to evaluative feedback. It is possible to estimate the value of this function by considering the project to be a Markov chain, and sampling a set of observation $\{ y \}$ from the process.

There are n levels of treatment Φ, including the stable value $\Phi_\circ$; $\Phi = \{ \Phi_1, \Phi_2 \dots \Phi_{\circ-1}, \Phi_\circ, \Phi_{\circ+1} \dots \Phi n \}$. Let $[p_{ij}]$ be a stochastic matrix, $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$, $(i, j = 1, 2, \dots n)$. Thus the state space is finite and the process is a Markov chain. The transition probability of moving from treatment Φ_i to Φ_j is p_{ij}. The probability $p_i{}^\circ$ gives the probability of finding the process initially at treatment level Φ_i. This can be interpreted as the point at which feedback occurs.

The transition probability p_{ij} is a function of an unknown parameter R. Let $p_{ij}(R) = R_{ij}$, for $\Phi_j = R\Phi_i$. We propose a test of the hypothesis that the stationary or terminal distribution of p_j is a specified set $\{ p_1, p_2, \dots p_n \}$, under the conditions that

$$\sum_i^n p_i \, p_{ij} = p_j \quad (j = 1, 2 \dots n).$$

By Theorem 5. 4 of Billingsley (2), we derive the equation set

$$\sum_j p_i \, p_{ij} = p_j \;, \quad p_{ij} = \frac{m_{ij}}{p_i \, \lambda_j + \mu_i}$$

with m_{ij} the transition frequences, λ and μ the counting measure on the state space. For $n > 2$, the equations must be solved iteratively. If q_{ij} is the solution, then the log likelihood statistic

$$\sum_{ij} m_{ij} \, L \frac{m_{ij}}{m_i \; q_{ij}} = \chi^2$$

asymptotically, with n-1 degrees of freedom (2: 64-65).

This statistic is appropriate for testing the hypot esis on p, and if

$$\sum \frac{(p_{ij} - q_{ij})^2}{q_{ij}} < \chi^2_{(\alpha, \, n-1)}$$

we suppose at a level of confidence α, that the stationary distribution of p is the specified set. This permits the estimation of R, as for each unique probability distribution p there is one reaction function R

REFERENCES

1. Welty, Gordon A., "The Use of Experimental Designs in the Decision-Making Feedback Process," *Journal of Experimental Education,* 37:2 p. 89, 1968; also cf. "Experimental Designs and Applied Research," *California Journa of Educational Research,* 20:1, p. 43, 1969.
2. Billingsley, Patrick, *Statistical Inference for Markov Processes*, University of Chicago Press, Chicago, 1961.